Computing expert's intelligence

Neocleous, Andreas

university of groningen

UNIVERSITY OF GRONINGEN

JOHANN BERNOULLI INSTITUTE FOR MATHEMATICS AND COMPUTER SCIENCE

COMPUTING EXPERT'S INTELLIGENCE: A CASE IN BIO-MEDICINE AND A CASE IN MUSICOLOGY

A dissertation supervised by promotors

PROF. DR. SC. TECHN. NICOLAI PETKOV

AND PROF. DR. CHRISTOS N. SCHIZAS

and submitted by

ANDREAS NEOCLEOUS

in fulfillment of the requirements for the Degree of

PHILOSOPHIÆDOCTOR (PH.D.)

December 2016

ISBN: 978-90-367-9393-3 (ISBN ebook: 978-90-367-9392-6)

university of groningen

Computing Expert's Intelligence: A Case in Bio-medicine and A Case in Musicology

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Tuesday 13 December 2016 at 9.00 hours

by

Andreas Neocleous

born on 18 July 1984
in Larnaca, Cyprus

Supervisors
Prof. N. Petkov
Prof. C. N. Schizas

Co-supervisor
Dr. G. Azzopardi

Assessment committee
Prof. M. Vento
Prof. C. Pattichis
Prof. M. Biehl
Prof. A. Telea

The research work presented in this thesis was carried out at the Intelligent Systems Group of the Johann Bernoulli Institute for Mathematics and Computer Science of the University of Groningen and the Intelligent Systems Group of the Computer Science Department of the University of Cyprus. It has been funded by the University of Groningen through an "Ubbo Emmius" scholarship for international sandwich PhD programs, the University of Cyprus, the Republic of Cyprus through the Cyprus Research Promotion Foundation and the Fetal Medicine Foundation of the UK.

Front cover: A fetus surrounded by some type of waves. It represents the moment of the ultrasonographic screening. The waves also represent the sound and the music.

ISBN: 978-90-367-9393-3 (printed version)
ISBN: 978-90-367-9392-6 (digital version)

Abstract

This thesis is a result of a sandwich PhD that has been done in collaboration between the University of Groningen and the University of Cyprus. The research work is presented here in two parts. The one comprises research on the applicability of machine learning techniques for the early identification of chromosomal abnormalities. The second part is focused on the application of these techniques on folk music identification and recognition. The ultimate objective of this thesis is the development of machine learning techniques that can be validated in real data in medicine and musicology and that can have practical value.

Through the collaboration with the Fetal Medicine Foundation (FMF), the potential use of artificial intelligence techniques for the prognosis of fetal chromosomal abnormalities in the first trimester of pregnancy is explored and reported in part I. A dataset of more than 100000 cases of pregnant women that underwent in a prenatal screening was used to train Artificial Neural Networks (ANNs). This dataset contains data for the normal cases (euploid) and five chromosomal abnormalities: Trisomies 21 (T21), 18 and 13, triploidy and Turner syndrome. Standard cross validation techniques had been applied to test the performance and the generalization abilities of the ANNs with unknown validation sets. It is shown in part I that the ANNs achieve better results than other existing methods in terms of diagnostic rate (DR) of chromosomal abnormalities (100% DR of T21) at a lower false positive rate.

The work presented in part II of the thesis has been done under a three-years research project for the analysis of the folk music of Cyprus and the Eastern Mediterranean Countries and it was funded by the Research Promotion Foundation of the Republic of Cyprus. The COSFIRE filters that have been found effective for 2D and 3D signals, had been adapted for 1D music signals and their effectiveness in different applications has been studied and the results are reported. More specifically, in Chapters 5 and 6 we use the COSFIRE filters for melodic similarity and ornamentation detection. We show that the proposed method using COSFIRE filters outperforms state-of-the-art methods such as the dynamic time warping that are used for similar tasks.

Samenvatting

Deze scriptie is het resultaat van een sandwich PhD die gedaan is in samenwerking met de universiteit van Groningen en de universiteit van Cyprus. Het onderzoekswerk wordt hier gepresenteerd in twee delen. Het ene deel bevat onderzoek naar de toepasbaarheid van zelflerende technieken voor de vroege identificatie van chromosomale afwijkingen. Het tweede deel richt zich op de toepassing van deze technieken op volksmuziek identificatie en herkenning. Uiteindelijk doel van deze scriptie is de ontwikkeling van zelflerende technieken die in echte data gevalideerd kunnen worden in medicijnen en musicologie en die van praktische waarde kunnen zijn.

In samenwerking met de Fetal Medicine Foundation werd het potentiële gebruik van kunstmatige intelligentie technieken voor de prognose van foetale chromosomale afwijkingen in het eerste trimester van zwangerschap onderzocht en gerapporteerd in deel I. Een dataset voor meer dan 100.000 gevallen van zwangere vrouwen die een prenatale screening kregen, is gebruikt om Kunstmatige Neurale Netwerken (KNNet) te trainen. Deze dataset bevat data van de normale gevallen en vijf chromosomale afwijkingen: Trisomieën (T21), 18 en 13, triploïde en Turner syndroom. Standaard kruisvalidatietechnieken zijn ingezet om de prestatie- en generalisatiecapaciteiten van de KNNet te testen met onbekende validatiesets. In deel I wordt aangetoond dat de KNNet betere resultaten behalen dan andere bestaande methodes op het gebied van diagnostische nauwkeurigheid (DR) van chromosomale afwijkingen (100% DR of T21) tegen een lager vals-positief resultaat.

Het werk gepresenteerd in deel II van de scriptie is gedaan in een driejarig onderzoeksproject voor de analyse van de volksmuziek van Cyprus en de oostelijk mediterrane landen en is gefinancierd door de Research Promotion Foundation van de republiek Cyprus. De COSFIRE filters die effectief zijn gebleken voor 2D of 3D signalen, zijn aangepast voor 1D muzieksignalen. Hun effectiviteit in verschillende applicaties werd bestudeerd en de resultaten zijn gerapporteerd. Meer specifiek, in Hoofdstuk 5 en 6 gebruiken we de COSFIRE filters voor melodische overeenkomst en ornamentiek detectie. We tonen aan dat de voorgestelde methode, die gebruik maakt van COSFIRE filters, beter presteert dan geavanceerde innovatieve methodes zoals de dynamische tijdvervorming die gebruikt worden voor soortgelijke taken.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my supervisors Professor Nicolai Petkov and Professor Christos N. Schizas who guided me in this work. Thank you for the trust you placed on me to work with you towards the Doctoral degree. Along with your excellent guidance towards completing my work, you gave me valuable advices on important life issues. You also gave me the opportunity to work and collaborate with people that are experts in their fields at the highest level. I am deeply grateful for your patience, support, promotion and supervision.

I would especially like to thank Professor Kypros Nikolaides for giving me the opportunity to work with valuable real medical data.

I would also like to thank Professors Costas Pattichis, Andreas Spanias, Michael Wilkinson and Michael Biehl for their interest on my work, their attendance on my talks and their valuable feedbacks and remarks.

I am thankful to the Assessment Committee members of the University of Groningen, Professors Mario Vento, Costas Pattichis, Michael Biehl, and Alex Telea for reviewing my work and for their valuable comments and remarks.

From the Intelligent Systems Group of Johann Bernoulli Institute I am deeply thankful to my co-supervisor George Azzopadi for the excellent collaboration at work but also as a great friend. The people I "grew up" together all these years, Laura Fernandez Robles, Nicola, Manuel, Estefania, special thanks to Astone, Jiapan, Ugo and Jonas for being flatmates at certain periods of time, Tiziana, Charmaine, Ahmad, Fritz, Baris, Laura Fiorini, Miguel, Rafael, Danilo, Daniel, Renata, Fatimah, Petra, Doina, Sofia and Walter.

From the Intelligent Systems Group of the University of Cyprus, I would like to thank Maria, Eirini, Natasa, Savvas, Zinonas, my brother Kleanthis, professor Costas Neocleous, Professor Marios Avraamides and the two medical doctors, Anastasia and Elina. I am also thankful to Mr. Terlikkas and Maria Panteli for the collaboration on the analysis of the folk music of Cyprus and Dr. Argiro Syngelaki for the collaboration on the data preparation of the fetal chromosomal abnormalities.

I would also like to express my gratitude to the administrative staff for their

immediate responses regarding paper work and financial issues. Specifically, Esmee Elshof, Desiree Hansen, Ineke Shelhaas and Annette Korringa, Elena Kakos, Zacharias Aristides.

I would like to thank the very first group of people I met in Groningen, that we really had an amazing time: Milos, Anita, Mariagni, Gabija, Lenka, Tomas, Sven, Alex, Chris, Emilia, Octavio, Giorgos, Georgia Dimitriou, Georgia Savvidou, Pampos, Elena, Maria, Ellie, Sotiris, Nicolas, Poly, Skevi, and Andreas.

I would also like to thank my paranymphs Nicola Strisciuglio and Jiapan Guo for supporting me and accepting to be next to me during my defence.

My gratitude is also conveyed to my family, my parents Costas and Yioula, my brother Kleanthis and his family (Eleni, Sofia and Melia) and my sister Barbara for their endless support, love and belief. Without you and your support, this life experience would be impossible.

Finally, I appreciate the financial support from Research Promotion Foundation, Fetal Medicine Foundation, Republic of Cyprus, University of Cyprus and University of Groningen that funded parts of the research discussed in this dissertation.

Chapter 1

Introduction

Classification and data analysis techniques are nowadays applied in most of the fields of science. In medicine, a significant amount of work has been done for predicting or diagnosing diseases, using historical data from examinations of previous patients (Fang et al. 2016).

For applications and tasks such as medical diagnosis, weather forecast, forex prediction, pattern recognition in arts and culture, speech recognition, text identification and others, typically the information is digitally stored in databases and processed computationally. In computer science, these datasets are used to build systems that automatically perform actions and return results. In this thesis, I used two different types of datasets to apply machine learning and signal processing techniques for two tasks. In the first case which involves the early detection of chromosomal abnormalities, the data that I worked on contain medical observations from pre-natal examinations, including both fetal and maternal characteristics. In the second case, signal processing techniques and filter based methods for the identification of repeating patterns and ornamentations in ethnomusicological data are proposed. These are presented in the second part of this thesis.

In medicine, data storage and health record of patients has been a common practice for medical doctors since the age of Hippocrates. Medical information related to the evolution and the causes of a disease was stored. The modern health record is structured in a more detailed way, including information from examinations such as blood tests, heart rate, allergies, biophysical measurements, medicine taken and many more. The advance of digital technology gave from one hand the opportunity to develop specialized equipment for medical examinations such as ultrasound screenings, MRI, automated systems for medicine control of patients, tools for surgeries etc, and on the other hand gave the ability of storing digital information from many private doctors, networks of hospitals from several countries operating electronic health record systems.

The benefits of using an electronic health record system that ideally will be globally interoperable are many, both on a personal level, as well as on a social level. Some of the benefits from the patients' point of view are the opportunities for a better health care, the safety of the personal data, the freedom of choosing more than one doctor or hospital for a diagnosis and treatment among others. On a social level,

the hospitals have digital libraries for the records of patients and there is substantially reduced need for manpower, papers and storage rooms. There is also better communication between the doctors of different hospitals. From the computer science and machine learning points of view, we take advantage of the multitude of data that are stored in such electronic health records in order to create systems that identify and predict several diseases and thus for better health management.

What is important in machine learning, is to analyze with big data and extract the knowledge and the usefulness through this analysis. In the last years, there has been an attempt to analyze unstructured data and build models that are able to identify patterns of interest. The Watson platform from IBM (Wolf et al. 2016) is an example of such systems. In medicine, such unstructured data can be for example the documents of the doctor.

The dataset I used for the identification of chromosomal abnormalities consists of normal cases, called euploid, and five abnormal cases: trisomies 21 (T21), 18 (T18) and 13 (T13), Turner syndrome and triploidy. More than 99% of the cases are euploid and thus it is a highly imbalanced set. It is common in medical datasets that the populations of the normal and the abnormal cases have an imbalanced relation due to the fact that from the entire population of cases that are examined, only a small percentage will be positive to a disease. Examples of such datasets among others, are described in (Klement et al. 2012) for computed tomography imaging of children, in (Boughorbel et al. 2016) for breast cancer prognosis, in (Sheikhi and Altınçay 2016) for type II diabetes detection and in (Dinov et al. 2016), for Parkinson's disease prediction.

The fact that in medical datasets typically the positive cases consist the minority of the total population is encouraging for human nature and health, however it is many times an obstacle for classification methods in terms of accurate generalizations. Every popular classifier such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), k-Nearest Neighbours (k-NN) and others have some drawbacks in their methodology regarding the class imbalanced problem. In the literature, a lot of work has been done for creating balanced from imbalanced datasets for classification (Mazurowski et al. 2008, Wilk et al. 2016, Hoens et al. 2013, Stefanowski and Wilk 2008). This is typically done either by oversampling the minority class (Pérez-Ortiz et al. 2015) or downsampling the majority class (Kubat et al. 1997).

In both approaches, it is important to make sure that the information that is stored in the data that are used for training is not lost when downsampling the majority class or distorted when oversampling the minority class. One significant contribution among others in this thesis, is the application of a method for downsampling the majority class using the k-means algorithm. It is based on the assumption that the euploid class can be represented with a number of subclasses. This assumption rises from the large amount of cases that consist the euploid class and thus the normality can be consisted with more than one pattern and it is difficult to be de-

fined. It is also possible that subclasses exist in the minority class of the aneuploidy cases, but it is difficult to safely drive such conclusion due to the limited number of cases. Nevertheless, if the above assumption is valid, then the supervised methods are forced to generalize a multi class feature space such as the euploid, into a single class. In Chapter 3, it is shown that ANNs perform better when the populations in both classes that are used for training are consisted with equal or similar number of cases.

The findings in this thesis contributes also to the determination of the relative significance that each feature has to the prediction or the diagnostic findings. This has been facilitated through heuristic and through formal mathematical approaches. In Chapter 3, I present the results of the ANNs that are built using different combinations of input markers. In (Neocleous et al. 2016b), we applied the Generalized Matrix Learning Vector Quantization (Biehl et al. 2006b) (GMLVQ) method to measure the relevances of the input markers. The results of the GMLVQ are in line with the ones presented in Chapters 2 and 3.

Concerning the data in ethnomusicology, these are consisted of one for monophonic or two for stereophonic dimensional feature spaces. These are typically the sound pressure over time for the two channels left and right, that are converted into electronic signal and stored in digital/audio format. Identifying specific features that model and distinguish the folk music of non-Western countries requires several adaptations of the existing algorithms in order to meet the particularities of each country's musical culture. Typically, for problems such as instrument identification, segmentation of important musical parts in a song and music similarity, a set of low features, called descriptors are extracted directly from the audio signal (Neocleous et al. 2014b). Then, by applying dynamic programming, higher level features such as the scale of a song can be extracted and used for classification and similarity measures (Gómez 2006). The contribution of this thesis in the field of the Computational Ethnomusicology is presented in Chapters 5 and 6. An adaptation of the COSFIRE filters (Azzopardi and Petkov 2013, Azzopardi and Petkov 2014) to 1D signals is proposed for tasks such as repeating pattern identification that is found in melodies and for ornamentation detection. Further work was done for symbolic representation of musical signals, tonal similarity, segmentation into important musical parts of songs and published in several conferences and workshops (Neocleous et al. 2012, Neocleous et al. 2014a, Neocleous et al. 2014b). The COSFIRE filters outperform state-of-the-art methods in digital signal processing such as dynamic time warping (DTW), as shown in Chapter 5.

1.1 Scope

In part I of this thesis, the problem of the early diagnosis of fetal chromosomal abnormalities during pregnancy is addressed. The goal of this study is to explore the potential use of machine learning techniques, such as ANNs, SVMs and k-NNs, in order to generate classification models for predicting fetal chromosomal abnormalities in the first trimester of pregnancy. The aim of the work done as presented in Chapters 2, 3 and 4 is to achieve 100% diagnostic rate of the T21 at the lowest false positive rate (FPR) possible.

Other equally important objective of this study include the optimization of a cut-off value (threshold) that is applied to the output of the ANN of an unknown case in order to binarize that value into a class between normal or abnormal. In most of the classification methods, there is a need of using a cut-off value that separates a similarity measure or a system output into a binary decision and that is in the range between 0 and 1. In the identification of chromosomal abnormalities, we represent the normal cases (euploid) with the number 0 and the abnormal cases with the number 1. A cut-off value that is closer to 0, returns high detection rates, at a cost of high FPR.

The question here is how important is to identify the entire population of the abnormal cases with respect to the number of the false positive classifications. These are cases that will have to perform additional unnecessary examinations, including invasive tests that may put a mother in risks for complications and the fetus in risks for miscarriage.

In the optimization stage of the cut-off value, we consider several social and financial aspects, such as the cost of a false positive classification with respect to the cost of a false negative classification. There are several methods to optimize the cut-off value as it has been discussed in (Neocleous et al. 2016). However, our criterion is to use a cut-off value that will return a 100% detection rate (DR) for T21.

In addition to the objectives stated above, several other research questions of equal importance are addressed that are related to practical aspects of the pre-natal examination procedure. We attempted to minimize the number of parameters that are needed as input to the system and at the same time to meet the above mentioned objective (100% DR for T21). In Chapter 3 we separate the available medical markers into two groups that come from two different examinations. From the findings of this study, we raised another question: is it possible to propose a two-stage screening for reducing the number of unnecessary examinations? From the findings in Chapter 4 it is shown that a two-stage approach can reduce dramatically the number of invasive tests that are currently performed. The research questions are summarized as follows:

- How can we build an effective system that will require the least possible prenatal examinations?
- Which machine learning techniques are appropriate for the modelling of medical data?
- How to deal with highly imbalanced datasets?
- How do ANNs contribute to the problem of the detection of chromosomal abnormalities?
- How to determine the best cut-off value that guarantees a 100% DR of T21?

In the second part of this thesis, I propose a novel algorithm for the identification of important patterns in 1D signals. This method is an adaptation of the COSFIRE filters that are found effective in 2D and 3D signals such as in image processing. The research questions in this part of the thesis are the following:

- Are the proposed COSFIRE filters effective and appropriate for applications using 1D signals?
- How does their performance compare with other state-of-the-art methods?
- What differentiates the COSFIRE filters from other existing methods?
- How can the COSFIRE filters be used in other 1D signals than the ethnomusicological data?

In Chapters 5 and 6 I first describe the proposed method and then I apply such filters to ethnomusicological data for the identification of repeating patterns and ornamentations. Particularly, in Chapter 5, I have applied the COSFIRE filters to a dataset of 38 songs that I created and published online. From the results, it is shown that COSFIRE filters outperform state-of-the-art methods. The COSFIRE filters are found to be effective in the identification of patterns in signals that have temporal variations, such as the singing voice. Similar to this idea is the well known method DTW (Müller 2007). However, what differentiate COSFIRE filters from DTW is the fact that the COSFIRE filters allow temporal tolerance that can be optimized with parameters. Additionally, COSFIRE filters are more computationally efficient than other methods. The COSFIRE filters proposed in this thesis can be applied in any type of 1D signal. However, additional experiments need to be done to test their performance on other tasks, such as the identification of k-complexes in EEG signals.

1.2 Outline

The rest of the thesis is based on journal and conference papers that are either published or accepted for publication or submitted to academic journals of high impact factors. Some of the content in this thesis might be repeated in order to make the chapters self contained.

In Chapter 2, the problem of the identification of chromosomal abnormalities is introduced. In that Chapter, a two-stage approach for the detection of the T21 (stage 1) and the detection of the other chromosomal abnormalities (stage 2) is proposed. A statistical analysis of the dataset is also presented and discussed with respect to the discriminative power of every independent variable.

In Chapter 3, a more technical approach to the problem is introduced. In that Chapter, an attempt is made to address questions such as whether to use normalized or raw data for training the classifiers, and on whether a balanced training set improves the performance of the models considering that the dataset is highly imbalanced, like in many medical datasets. Finally, experiments have been done in a heuristic way to explore the contribution of the parameters to the system.

In Chapter 4, a two-stage procedure is proposed for the detection of the chromosomal abnormalities. This is an extension of the work that is presented in Chapter 5. It is based on the premise that pregnant women perform the standard screening in the first trimester of pregnancy which contains an ultrasound examination and a blood test. In the first stage of our trials, when using 4 parameters (maternal age, PAPP-A, β-hCG and the nuchal translucency), we achieve 100% DR of T21 at a relatively high FPR of about 20%. In the second stage, all the cases that are ranked as positive in stage 1, are further processed with another examination which includes a specific ultrasonoghraphic scan to get information about the flow in two arteries of the fetal cardio system. In stage 2, we achieve 100% DR of T21 and more than 80% of the other chromosomal abnormalities, at a FPR of less than 3%.

In Chapter 5, we apply the COSFIRE filters for the identification of repeating patterns in ethnomusicological data. We have adapted the COSFIRE filters that are found effective in 2D and 3D signals into 1D signals. It is shown in that Chapter that the COSFIRE filters are effective for such problems and they can be applied in signals of different scales.

In Chapter 6, I introduce another application for ornamentation detection in singing folk music that is approached with the use of COSFIRE filters. It is shown that the COSFIRE filters are able to capture such particular variations in 1D signals and to identify similar patterns.

Conclusions and future work of this thesis are reported in Chapter 7.

Part I

Computational Intelligence in Medicine

Published as:

A. Neocleous, K. Nicolaides, and C. Schizas, "First Trimester Non-invasive Prenatal Diagnosis: A Computational Intelligence Approach," *IEEE Journal of Biomedical and Health Informatics*, Vol. 20, 2015.

Chapter 2

First Trimester Non-invasive Prenatal Diagnosis of Chromosomal Abnormalities

Abstract

The objective of the current study is to examine the potential value of using machine learning techniques such as artificial neural network (ANN) schemes for the non-invasive estimation, at 11-13 weeks of gestation, the risk for euploidy, trisomy 21 (T21) and other chromosomal aneuploidies (OCA), from suitable sonographic, biochemical markers, and other relevant data. A database[1] consisted of 51208 singleton pregnancy cases while undergoing first trimester screening for aneuploidies has been used for the building, training and verification of the proposed method. From all the data collected for each case from the mother and the fetus, the following nine are considered by the collaborating obstetricians as the most relevant to the problem in question: maternal age, previous pregnancy with T21, fetal crown-rump length, serum free β-hCG in multiples of the median (MoM), PAPP-A in MoM, nuchal translucency thickness, nasal bone, tricuspid flow and ductus venosus flow. The dataset was randomly divided into a training set that was used to guide the development of various ANN schemes, support vector machines and k-nearest neighbours models. An evaluation set used to determine the performance of the developed systems. The evaluation set, totally unknown to the proposed system contained 16,898 cases of euploidy fetuses, 129 cases of T21 and 76 cases of OCA The best results were obtained by the ANN system which identified correctly all T21 cases i.e. 0% false negative rate (FNR) and 96.1% of euploidies i.e. 3.9% false positive rate (FPR), meaning that no child would have been born unexpectedly with T21 if only that 3.9% of all pregnancies had been sent for invasive testing. The aim of this work is to produce a practical tool for the obstetrician which will ideally provide 0% FNR, and recommend the minimum possible number of cases for further testing such as invasive. In conclusion it was demonstrated that ANN schemes can provide an effective early screening for fetal aneuploidies at a low FPR with results that compare favorably to those of existing systems.

[1]The data set can become available for academic purposes by communicating directly with the authors.

2.1 Introduction

The identification of chromosomal abnormalities in the early stages of pregnancy can be achieved, with high confidence, by performing an amniocentesis test or a chorionic villus sampling test (CVS) (Brambati et al. 1985). These methods however, are invasive and carry high risk for miscarriage or cause medical side effects to the pregnant woman.

2.1.1 Invasive Methods

In the work of Tabor and Alfirevic (Tabor and Alfirevic 2009), it is reported that amniocentesis and CVS have a miscarriage rate of 0.5 - 1.0%. Furthermore, the amniocentesis test should not be performed prior to the 15 weeks of pregnancy since the miscarriage rate increases significantly, while there is also increased risk of developing talipes equinovarus (Sundberg et al. 1997).

Therefore, the amniocentesis or the CVS test should not be performed unless there are serious indications of high risk for chromosomal aneuploidy in the fetus. It is also emphasized that both of these methods carry additional costs to the family which is not a trivial issue.

2.1.2 Non-Invasive Methods

There has been an increased interest and need of exploring non-invasive methods for the prediction of chromosomal aneuploidies in the first trimester, or earlier, of pregnancy. In the literature, several methods that were aiming at a non-invasive prediction of chromosomal abnormalities had been reported. These methods can be classified into three main categories, based on their methodology:

First-trimester pre-natal statistical screening

Statistical methods for appraising the probability of aneuploidy are properly applied on several markers that are obtained by an antenatal test. Typically, the markers used are the fetus crown ramp length (CRL), the fetal nuchal translucency (NT), the maternal age (MA), the pregnancy-associated plasma protein-A (PAPP-A), the serum free β-hCG, the ductus venosus flow (DV), tricuspid valve flow and others (Nicolaides 2011, Spencer 2007, Dugoff et al. 2010). The risk for aneuploidies increases with maternal age and it is higher in women with previous affected pregnancies. It also increases with fetal nuchal translucency thickness and is higher in those with absence of the fetal nasal bone and with abnormal flow through the ductus venosus and across the tricuspid valve. The maternal serum concentration of the placental products free β-human chorionic gonadotropin and pregnancy associated plasma protein-A also influence the risk for aneuploidy (Nicolaides 2011). Serum

PAPP-A is decreased in trisomies 21, 18, 13 and the Turner syndrome, while serum free β-hCG is increased in T21, decreased in trisomies 18 and 13 and not altered in the Turner syndrome.

Most of these methods use posterior probabilities based on the median and the standard deviation of the markers, or by using a suitable multivariate statistical approach. Typically, the models output a probability on the risk of fetus aneuploidy. In the work of Nicolaides et al. (Nicolaides et al. 2005), a multivariate likelihood approach is described for the identification of the T21 in the first trimester of pregnancy. A multiplication of the maternal age-related risk and each likelihood ratio (LR) derived from the fetal NT and maternal weight-adjusted serum free β-hCG and PAPP-A outputs the patient-specific risks for aneuploidy. More on this method can be found in (Reynolds and Penney 1990).

Traditionally, screening for aneuploidies is done by estimating the patient-specific risk for each aneuploidy. This is done by multiplying the a priori risk, by a factor of approximately 1.75 in cases with a previous history of aneuploidy. The likelihood ratio for each ultrasonographic and biochemical marker is used to calculate the detection and FPRs by taking the proportions with risks above a given risk threshold (Nicolaides 2011). The likelihood ratios for the categorical variables such as the absence or presence of nasal bone and the abnormality of flow in the tricuspid or ductus venosus valves, are ratios of prevalence of each marker in each type of fetal aneuploidy to the prevalence in euploidy fetuses. In the case of continuous variables, such as fetal NT thickness and maternal serum free β-hCG and PAPP-A, the likelihood ratios are derived from the overlap of the Gaussian distribution of each marker in each type of fetal aneuploidy with the respective Gaussian distribution in euploidy fetuses.

Maternal cell-free DNA screening

Schmorl's experiment with women who died of eclampsia was the first study to suggest that fetal cells circulate in the mother's blood (Schmorl 1893). It is estimated that 1 in 1000 to 1 in 10000000 nucleated cells in maternal blood are fetal (Bianchi et al. 1990). The isolation of fetal DNA from the maternal DNA could give insights for studying genetic diseases. For the problem under study, the isolation of the fetus DNA is still extremely hard to achieve, it is labor intensive and requires highly skilled operators. While the majority of the studies focused on the identification of the T21 (Ehrich et al. 2011, Chiu et al. 2011), recently the trisomies 18 and 13 are also identified. In (Papageorgiou et al. 2011), Papageorgiou et al. claim a very good prediction (100%) of 26 euploidy and 14 cases of T21 in the validation set, by determining the DNA methylation ratio of 12 selected differentially methylated regions (DMRs).

Other studies used similar approaches, mainly for T21, trisomy 18 (T18) and T13 (T13) (Myers et al. 2004, Ashoor et al. 2012, Palomaki et al. 2011). Palomaki et al., (Palomaki et al. 2011) report a detection rate of 100% for T18 and 91.7% for T13 with a FPR of 0.28% and 0.97% respectively. However, the test was done in the late first and early second trimester of the pregnancy which can be considered relatively late for abortion. Most importantly, the circulating cell-free DNA fragments are being presented as differentially methylated markers and not identified under microscope. Then, the procedure for the prediction of the risk for aneuploidies is done with simple statistical analysis of the methylated markers.

The above described non-invasive methodologies have their relative advantages but at the same time have their disadvantages. The first methodology suffers in the sense that one cannot combine and examine simultaneously all the relevant parameters of the case or visualize them in a multidimensional space. Visualization of more than three parameters at a time is extremely difficult and not practical in a medical environment. Also, parameters which are correlated can lead to unreliable conclusions. In the second methodology, the parameters used (called markers) have no relation to the phenotype of the fetus, such as for instance the CRL or the fetal heart rate which for the certain problem is important. In the majority of the published studies the population of their databases is very small for drawing reliable conclusions for such a complex problem. For instance in (Papageorgiou et al. 2011), they are dealing with 40 cases (euploidy and T21). Even though this work was published in Nature Medicine, one can identify two important limitations of the proposed method which imply lack of scientific confirmation. For such a complex problem, it is not convincing whether a perfect prediction of a population of 40 cases is satisfactory for drawing robust conclusions. In contrast to this database, we studied more than 51000 cases of pregnant women and validated 16,898 cases of euploidy fetuses, 129 cases of T21 and 76 cases of OCA. Also, there is no information in (Papageorgiou et al. 2011) whether the results were cross validated. It is important to present cross validated results such as 10-fold or leave-one-out cross validation. In this scenario, one can have a better insight about the generalization ability of the method. Therefore, it is not explicit if their method will yield similar results by randomly selecting different training and test sets. Another serious limitation of the study (Papageorgiou et al. 2011) is the fact that only euploidy and T21 can be handled; what will the system predict if the unknown case in question is T13 or OCA? Furthermore, the whole analysis even though is based on the existence of free DNA cells of the fetus in the mother's blood no such cell is isolated for singling out the pathological gene. The determination of a pathological existence in the genes is done statistically and probabilistically.

Computational Intelligence

In this study, we report a computational intelligent approach for the non-invasive estimation of the risk of aneuploidies. This approach involves the development of a system-predictor which takes as input a number of parameter values. These values have different origins and source, and they are collected at certain pre-specified times during pregnancy. For example during the first-trimester screening for fetal T21 and for OCA, certain parameter values are recorded which are a combination of maternal and feto-placental nature (Nicolaides 2011).

The computational ANNs, which are a specific paradigm of computational intelligence, had been used as effective classifiers and predictors for the last 25 years. Indeed, they had been extensively applied in medical and biological research and applications (Patel and Goyal 2007, Schnorrenberg et al. 1997, Schizas and Pattichis 1997, Neocleous et al. 2010).

ANNs are essentially mathematical algorithms implemented in software that learn from data and capture the knowledge and the internal dynamics that are contained in the data. Suitably trained models approach the functionality of small biological neural systems in a very fundamental manner that mimics human-like behavior. Thus, once they are properly trained they exhibit computational intelligence in a simplistic mimicry of the biological intelligence. They constitute a very simple digitized model of the biological brain and in some cases can detect complex non-linear relationships between dependent and independent variables in a dataset which may be undetectable by a human brain. Indeed, they can execute certain tasks, especially in classification and recognition that would be extremely difficult to be done by the traditional and conventional computing techniques. They can learn from data, even in self-organized manners.

In the medical field, ANNs proved to be a powerful method for medical diagnosis. As an example, Al-Shayea (Al-Shayea 2011) reports a medical diagnosis system for acute nephritis disease and heart disease using feed forward ANN. A correct classification of 99% has been reported. Hayashi and Setiono (Hayashi and Setiono 2002) used a two level approach combining two ANN systems for the diagnosis of hepatobiliary disorders. In this study, the database consisted by 536 samples with nine input features describing four hepatobiliary disorders: Alcoholic Liver Disease (ALD), Primary Hepatoma (PH), Liver Cirrhosis (LC) and Cholelithiasis (C). Their best ANN models classified 95% of the four diseases. A chest disease diagnosis system is reported in (Er et al. 2010). The database in this work consisted of chest disease measurements of 357 samples and six classes, namely Tuberculosis, COPD, Pneumonia, Asthma, Lung Cancer and Normal. All samples had thirty eight features. The authors report 90.2% average classification accuracy for all the six classes. Other studies that used ANN are referenced in (Khan et al. 2013, Maithili et al. 2011) and (Stephan et al. 2008).

The objective of our study is to examine the potential value of ANNs and other computational intelligence techniques in the prediction of the risk for T21 and OCA from ultrasonographic and biochemical markers at 11-13 weeks of gestation.

In section 2.2, computational intelligence approach and the proposed method are presented and discussed. Furthermore, statistical analysis has been applied to the non-binary features. This analysis is presented including the visualization of the feature distributions, and the testing of the separability of the features in pairs. In section 2.3 we present our experiments and results and in section 2.4 we discuss further the results of the present work and how it is compared with other methods. Finally, in section 2.5 we report our conclusions.

2.2 Methods

In this section we present the computational intelligence approach that has been adopted, the procedure for data collection, the data grouping into cross-validation sets, and two schemes for aneuploidy risk prediction. We have implemented models with ANNs, support vector machines (SVM) (kernel 1 and 2) and k-nearest neighbours (k-NN).

2.2.1 The artificial neural network diagnostic system

Many different ANN structures had been proposed and used by researchers in different fields. The most widely used ANN structure is the fully connected multilayer feed forward structure (FCMLFF). Mathematically this is represented by Eq. 2.1 as:

$$y_{iL}^{[L]} = f_{iL}^{[L]} \left(\sum_{j_{L-1}=1}^{n_{L-1}} y_{jL-1}^{[L-1]} W_{L-1,L}^{[L]} \right) \tag{2.1}$$

Where $y_{iL}^{[L]}$ is the output value of each neuron i of layer L that has a total of nL neurons. Typically, the function has a squashing form such as the logistic or the hyperbolic tangent. $W_{L-1,L}$ is the set of weights associating neurons in layer $L-1$ to neurons in layer L.

In this study, we used fully connected feed-forward ANNs as described in Eq. 4.2. The first layer is typically called input layer and it has as many neurons as the input parameters. The weights in the ANN represent the intensity of the processes in the synapses of the biological neural network. In a practical implementation of ANNs, the initial values of the weights are typically set to random values.

Once the ANN topology is decided, an effective training and tuning procedure needs to be implemented, so that the network will achieve the capability for generalization as a risk estimator. Many training procedures had been proposed and are

available for implementation. The most widely used for feed forward networks is the backpropagation algorithm (Werbos 1974). In this work, we implemented fully connected feed-forward neural networks with backpropagation learning. The justification for selecting this simple network is discussed further down. The activation function for the input layer and the hidden layer was set to logistic. We have trained and validated a large number of networks by changing different parameters in the training procedure. The best results were obtained with the networks built with 20 to 40 neurons in the hidden layer as will be explained in the results.

2.2.2 Database and Parameters Used

The data set used was collected from women having singleton pregnancies while attending the Fetal Medicine Centre at Kings' College Hospital and University College London Hospital in London, for aneuploidy screening at 11+0 to 13+6 weeks of gestation. The maternal age and the previous history of the pregnancy, in particular on whether there was a previous case of T21 were recorded. Also, a transabdominal ultrasound examination was performed for measurement of the fetal CRL and the NT thickness, as well as an assessment of the fetal nasal bone and the flow in the DV and across the tricuspid valve. These were done by sonographers who had received the appropriate Fetal Medicine Foundation Certificates of Competency. The pregnancy was dated according to the measurement of the fetal CRL (Robinson and Fleming 1975). Additionally, maternal blood was collected and used to measure serum PAPP-A and serum free β-hCG concentrations through automated machines that provide reproducible results within 30 min (Delfia Express System, Perkin Elmer). The measured PAPP-A and serum free β-hCG were converted into multiples of the median (MoM). The measured fetal NT was expressed as a difference from the expected normal mean for the specific CRL (Wright et al. 2008). These maternal demographic characteristics, ultrasonographic measurements and the biochemical results were recorded in a structured database.

The following nine parameters were used as suitable markers that could help in establishing the risk for aneuploidies: maternal age in years, history of previous pregnancy with T21, fetal CRL in mm, serum free β-hCG in MoM, PAPP-A in MoM, delta NT in mm, nasal bone (present or absent), tricuspid flow (regurgitation or normal) and ductus venosus flow (reversed a-wave or normal).

2.2.3 Statistical Analysis of the Data

In this section we present the results of a statistical analysis that has been applied to the data. The aim of this analysis is to discover the significance of separability between pairs of distributions for each feature, by means of their medians. That is, to test for the statistical null hypothesis whether data who belong to same category

share equal medians. Moreover, we have tested the normality of the distribution of each feature. This was done to create a better understanding of our data, as well as to avoid using methods that require normal distributions in the input feature space, such as the gaussian mixture models and the Student's t-test. The histogram of each non-binary feature has been computed for all the classes of our database, namely euploidy, T13, T21, T18, Triploidy and Turner. The normality of the data was tested with the Kolmogorov-Smirnov test (Kolomogoroff 2013).

In Table 2.1 the results of the normality test are shown. In the first row the names of the five non-binary markers are shown, while in the first column the names of the classes (euploidy or aneuploidies). It should have been expected that the data used in this study will not follow a normal distribution since the population under study is not normally distributed by nature, since pregnancy occurs in a certain age-range which is naturally skew to the right; in this large database used in this study contains proportionally much more trisomy cases than reported statistically. For example T21 occurs 1 in about 800 pregnancies, and thus in our database we should have had 65 cases instead of 408. Similarly, T18 occurs 1 in 5000 pregnancies and thus we should have had 10 instead of 145 cases, T13 1 in about 16000 pregnancies, etc.

The results of the normality test confirm this reality. Another factor which may had contributed to this is the fact that our database is highly imbalanced, ie. euploities are by far more than aneuploidies. In large sample sizes, the probability that a normality test will reject the null hypothesis that the sample comes from a normal distribution increases due to samples that depart from normality which are statistically significant. Furthermore, the population in the classes of T13, Triploidy, and Turner is relatively small (< 50) and therefore the rejection of the null hypothesis may be because of insufficient number of samples.

A widely used method to measure the separability between two classes for a given feature is the Student's t-test. The Student's t-test assumes that the testing data follow a normal distribution and the population of the two classes is equal. The t-test rejects the null hypothesis at a confidence level of 95% ($p < 0.05$) where p is the estimated probability of rejecting the null hypothesis.

Taking into consideration that a subset of the features does not follow a normal distribution and the population between the classes is not equal, the t-test could not be applied.

The significance of the difference between the medians of every class for the non-binary features was computed by the paired-difference Wilcoxon Rank Sum method (Wilcoxon 1945). The Wilcoxon Rank Sum method is a non-parametric method in contrast with the widely used t-test method. It has been proposed by Frank Wilcoxon in 1945 and popularized by Sidney Siegel in 1956 in applications to behavioural sciences. The p-values of the Wilcoxon Rank Sum method are summarized in Table 2.2. The pairs for testing the null hypothesis are shown in the first

Table 2.1: Results from the Kolmogorov-Smirnov test (Normality test). "MA" stands for maternal age, "N" stands for normal distribution while "NN" stands for non-normal distributions.

	MA	CRL	PAPP-A	β-hCG	delta NT
Euploidy	NN	NN	NN	NN	NN
T13	N	N	NN	N	N
T18	N	N	NN	NN	N
T21	NN	N	NN	NN	NN
Triploidy	N	N	NN	NN	NN
Turner	N	N	N	NN	N

column in Table 2.2. The word "True" signifies rejection of the null hypothesis and "False" acceptance. The null hypothesis that the medians of the data being compared have no statistical significance is rejected with a p-value that is less than 0.05. The method rejected the null hypothesis for all the features for the pairs "euploidy & T13", "euploidy & T18", "euploidy & T21, T18 & Turner", "T21 & Triploidy", and "T21 & Turner" and therefore the results were excluded from Table 2.2.

In Figs 2.1 to 2.5 we use the box-plots to present the median, the standard deviation, the normality of the distribution and the outliers of each feature. The total range of each feature is shown with a vertical dashed line while the lowest and the highest values are shown with small horizontal lines at its ends. The box in the middle represents the distribution of the feature in the range between the points of the first quartile (Q1) and the third quartile (Q3), where Q1 is defined as the middle number between the smallest number and the median of the data set. Similarly, Q3 is defined as the middle value between the median and the highest value of the data set. The median of each feature is shown with a small horizontal line within the box. In each figure, a box-plot of a feature is plotted against all the six classes. A dashed and a solid horizontal lines show the median of the euploidy and the median of the T21 respectively. These two lines were placed manually for visualization purposes.

The purpose of this analysis is to gain additional knowledge about the statistical properties and the significance of the separability of every feature. For instance, the medians of the populations of each class can be roughly compared by simply observing the distance between them. Also, it is interesting to visualize how the distribution around the median is spread. This can be done by observing the area enclosed in the box around the median. The normality of the distribution can also be observed if the median lies in the middle of the box. The outliers are also marked by a cross allowing a rapid way to estimate their population and their statistical

significance with respect to the distance from the median.

Fig. 2.1 shows statistical information of the fetal CRL parameter. The value of the CRL is extracted during the ultrasound screening. The doctor is manually annotating the preferred positions of the crown ramp by looking at the fetus on the screen. The manual annotation of the CRL creates a possibility for human errors in the measurement. Indeed, it may be that the high variance of the data for all the classes can be explained due to this observer error. It is worth mentioning that the

Table 2.2: The significance of the difference between the medians of every class computed by the paired-difference using the Wilcoxon Rank Sum method.

Pairs tested	MA	CRL	PAPP-A	β-hCG	delta NT
Euploidy	0.89	<<0.05	<<0.05	<<0.05	0.99
Triploidy	True	False	False	False	True
Euploidy	0.08	0.14	<<0.05	<<0.18	<<0.05
Turner	True	True	False	True	False
T13	0.14	0.28	0.02	<<0.05	0.08
T18	True	True	False	False	True
T13	<<0.05	<<0.05	<<0.05	<<0.05	0.06
T21	False	False	False	False	True
T13	0.04	0.3	0.03	0.05	<<0.05
Triploidy	False	True	True	True	False
T13	<<0.05	0.19	<<0.05	<<0.05	<<0.05
Turner	False	True	False	False	False
T18	0.13	<<0.05	<<0.05	<<0.05	<<0.05
T21	True	False	False	False	False
T18	<<0.05	0.79	0.07	0.35	<<0.05
Triploidy	False	True	True	True	False
Triploidy	0.32	0.04	<<0.05	0.03	<<0.05
Turner	True	False	False	False	False

median of the T21 is significantly higher from the euploidy while the median of the OCA is lower. This fact also explains that a system which in the training procedure takes all the trisomies as a single class may not achieve a robust generalization.

In Fig. 2.2 it is shown that the variance of the data for the serum free β-hCG feature is narrow with an exception on the data of the triploidy. As can be seen, there is one extreme outlier that forces the data to a non-gaussian distribution. This is reasonable, taking into account the small population of in this class. The feature serum free β-hCG has high separability by means of their median, for the pair "euploidy & T21" and low separability for the pairs "euploidy & Turner", "T13 & Triploidy" and "T18 & Triploidy".

The analysis of the feature Delta NT is shown in Fig. 2.3. The variance of the data in the classes T13, T18 and Turner is relatively higher with respect to the data in the euploidy and triploidy classes. It is also shown that the separability between the euploidy and the trisomies T13, T18, T21 and Turner has statistical significance.

It is well known that the maternal age plays a significant role for the classification of the euploidy or T21. This can also be observed in Fig. 2.4. The mean maternal age of the euploidy is 33 years while for the T21 is 38 years. It can be seen also that the means of euploidy and triploidy do not have significant statistical differences. While the mean of the materal age of T21, T13 and T18 is significantly higher,the mean for Turner is significantly lower. This can be seen in Fig. 2.4 and mathematically by the Wilcoxon Rank Sum method, shown in Table 2.2.

There is a statistical significance in the difference of the means between euploidy and the rest of the trisomies for the feature PAPP-A. In Fig. 2.5 it is shown that the mean of the euploidy has a higher value with respect to the means of the OCA The variance of all the classes for this feature is relatively low. In addition to the statistical analysis done in these data and the useful information regarding the nature of the data, an expert obstetrician can use these graphs while examining a new case. It can provide to him a tool for visualizing a new case with respect to thousands of previously confirmed cases.

2.2.4 Cross validation

The systems and approach described in this chapter were tested with a 3-fold cross validation. This was done by randomly dividing the data of 51208 (50517 euploidy, 408 T21, and 283 OCA) pregnancies into three training and evaluation sets containing proportionally the same numbers of euploidy and aneuploidy cases, as shown in Tables 2.3 and 2.4.

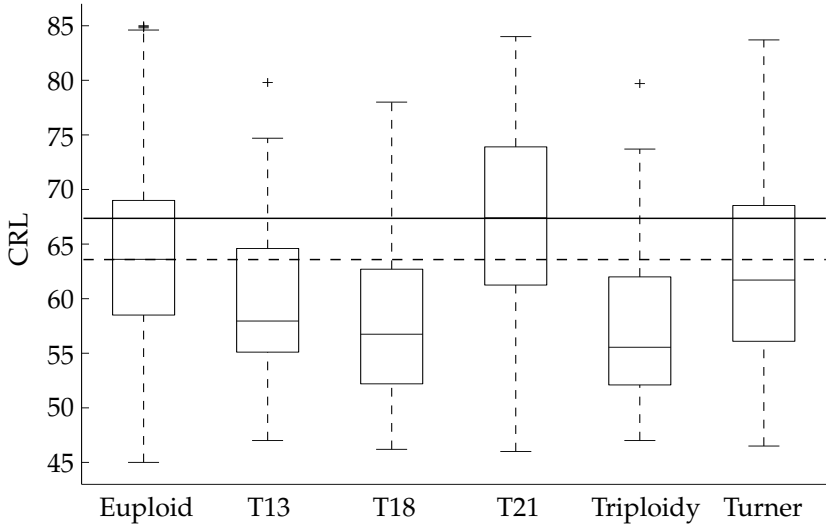

Figure 2.1: Visualization of statistical properties of the CRL feature.

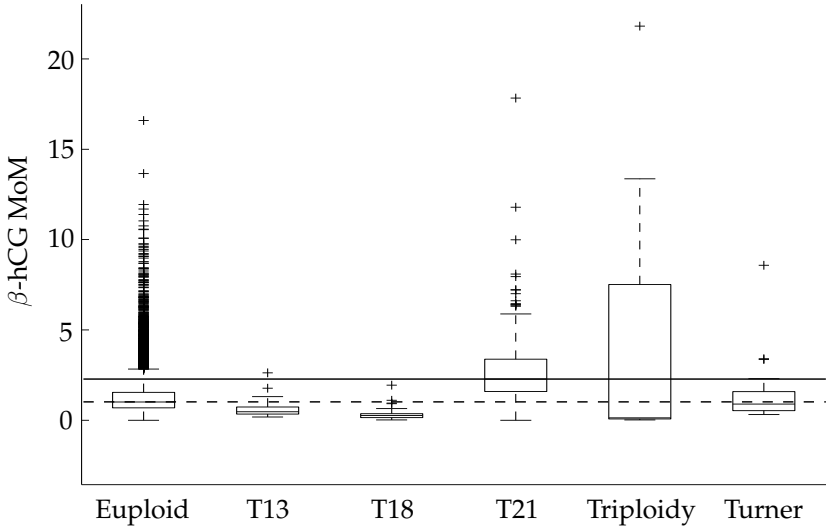

Figure 2.2: Visualization of statistical properties of the β-hCG MoM feature.

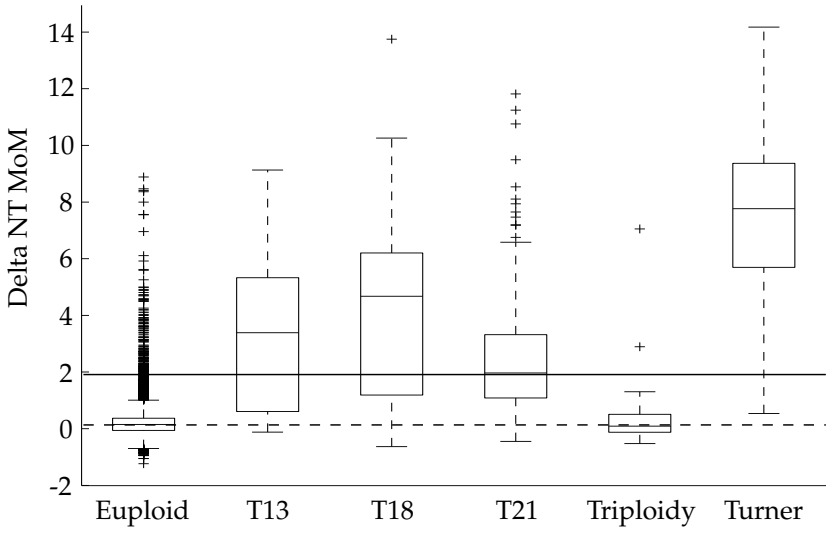

Figure 2.3: Visualization of statistical properties of the Delta NT feature.

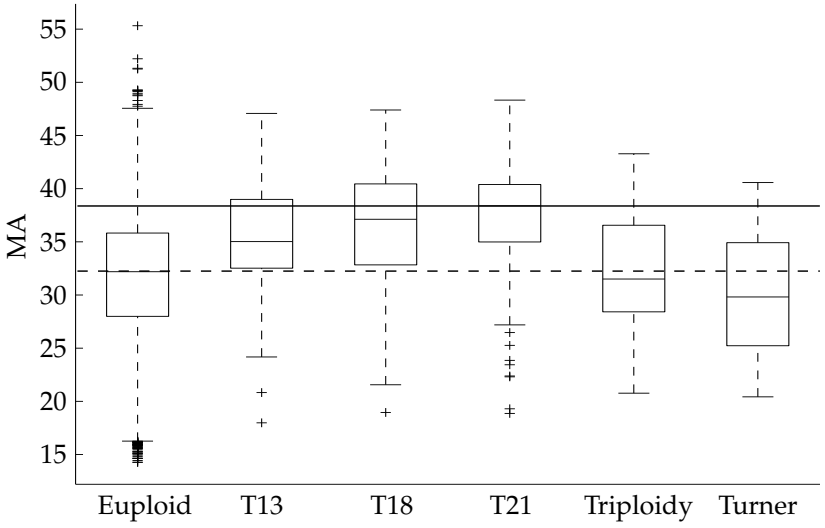

Figure 2.4: Visualization of statistical properties of the Mother's age feature.

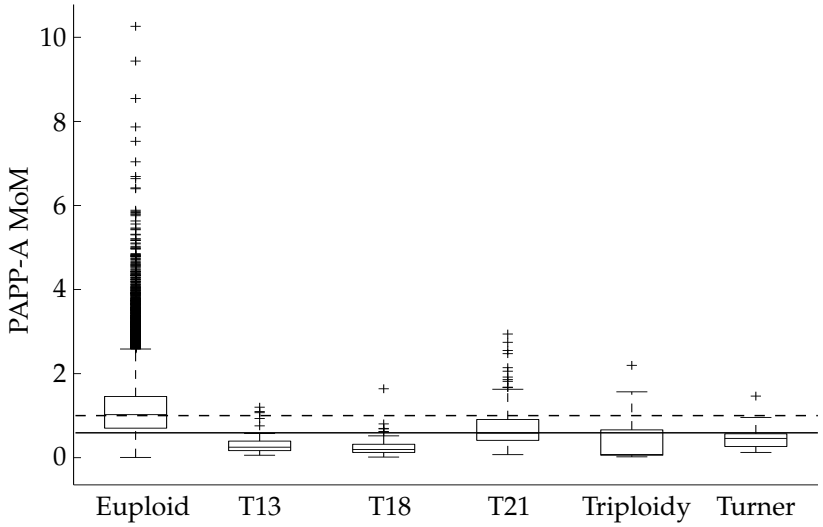

Figure 2.5: Visualization of statistical properties of the PAPP-A MoM feature.

Table 2.3: The number of cases that were used for training in the cross-validation procedure.

Training	Euploidy	T21	T18	T13	Triploidy	Turner
Fold 1	33,619	279	106	38	22	41
Fold 2	33,840	278	105	39	23	40
Fold 3	33,982	277	104	37	24	39

2.2.5 System 1: Classification into two classes: euploidy and T21

The dataset was split into a training and an evaluation set. The training dataset consisted only of euploidy and T21 pregnancies (Fig. 2.6), whereas the evaluation set contained also cases with OCA Various supervised models with ANN, SVM (kernel 1 and 2) and k-NN were developed and the best results in separating euploidy from T21 pregnancies were achieved by using a standard multilayer feed-forward neural network with one hidden layer. There were nine neurons in the input layer (9-input model), representing the nine markers that were used for training the networks (Table 2.5). The network output target was set to 0 for T21 and 1 for euploidy.

Table 2.4: The number of cases that were used for validation in the cross-validation procedure.

Validation	Euploidy	T21	T18	T13	Triploidy	Turner
Fold 1	16,898	129	39	14	10	13
Fold 2	16,677	130	40	13	9	14
Fold 3	16,535	131	41	15	8	15

After completion of the learning process through the use of the training dataset, the system was tested by the evaluation dataset which consisted of euploidy, T21 and OCA It is important to mention that system 1 can be considered as an autonomous system where T21 cases are detected with considerably low FPR. The drawback of this system is that the OCA are mostly predicted as euploidy. These false negative predictions of the OCA are important to be identified as abnormalities. While generally the OCA have an increased possibility of miscarriage during pregnancy, and therefore the early diagnosis of such abnormalities are not considered to have equal importance to the T21, in some cases the embryo survives until the late stages of pregnancy, or it is born and die few days later. This fact may cause health complications to the mother, and generate additional psychological damage to the relatives.

In order to predict correctly the OCA we propose system 2 that minimizes the FNR with a cost of increasing the FPR and causing a false alarm to some families. The doctors may use system 1 or system 2, having in mind the cost and the risk of sending a euploidy for further invasive examinations or considering an abnormal case as euploidy and let the pregnancy continue naturally.

2.2.6 System 2: Separate classification of three classes: euploidies, T21 and all the OCA

The same approach followed in developing system 1 had been followed for generating suitable classification models that could separate not only T21 from euploidy cases but also to test for the capability to separate the OCA from the euploidy cases.

In a first attempt we tried to build a neural network that could predict six situations. These were set as the outputs of the ANN. They were the euploidy, T21, T18, T13, Turner syndrome and triploidy. Although the cases of T21 were separable from euploidy, when the OCA were involved during the training phase of the network, this separation deteriorated and most of the OCA were classified as euploidy. Thus, this approach was abandoned.

A second approach that we tested was to build neural networks that were trained

Table 2.5: The parameters used as input vectors for the training models

Parameter	Model of 6 inputs	Model of 9 inputs
Maternal age	Used	Used
History of previous T21	Used	Used
Crown rump length	Used	Used
Delta nuchal translucency	Used	Used
Serum PAPP-A	Used	Used
Serum free β-hCG	Used	Used
Nasal bone	Not Used	Used
Ductus venosus flow	Not Used	Used
Tricuspid flow	Not Used	Used

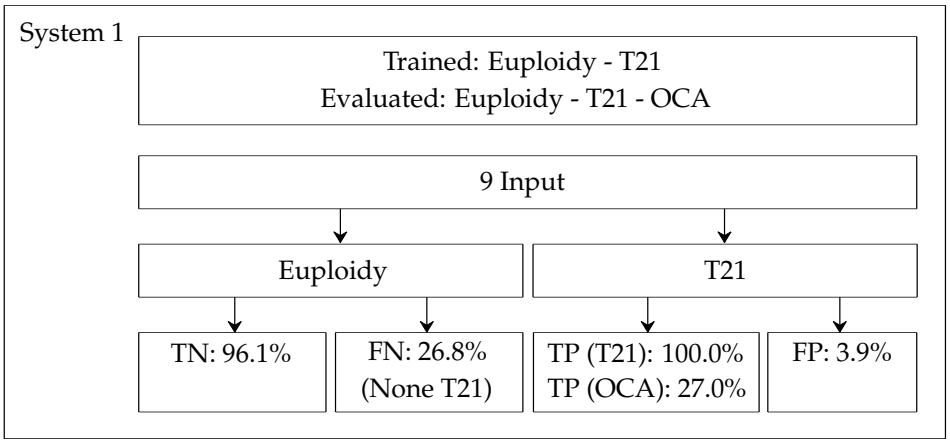

Figure 2.6: System 1, distinguishing between euploidy and T21 for the 9 input ANN. In this system the training set contained only euploidy and T21 cases. The evaluation set contained euploidy, T21 and OCA

to separate euploidy from T21, euploidy from T18, euploidy from T13, euploidy from Turner syndrome, euploidy from triploidy and T21 from OCA. This approach was also abandoned since it failed to successfully separate the groups. It was observed however, that most of the false positive cases (i.e. euploidy cases classified as aneuploidy) were the same for all the models examined. Thus, it was decided to exclude the models "euploidy & T13" and "euploidy & triploidy" from the overall

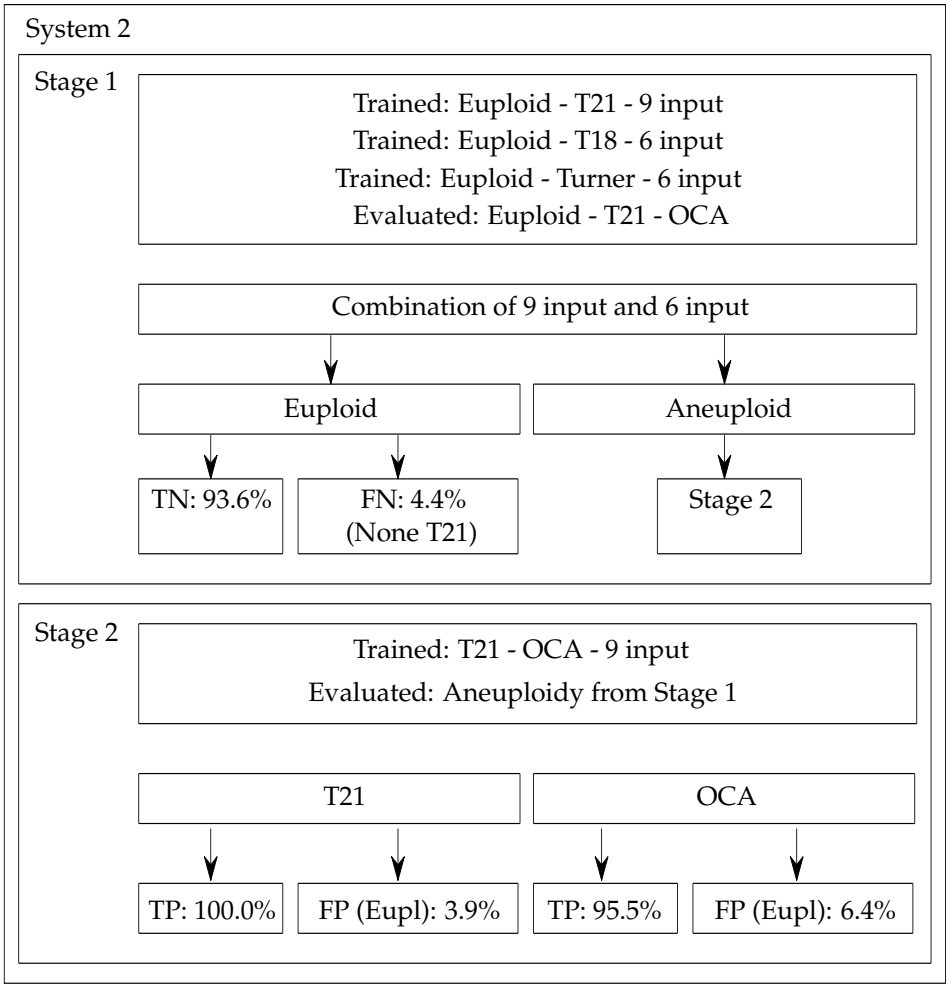

Figure 2.7: System 2, distinguishing between euploidy, T21 and OCA It is a combination of four ANN trained with 1) euploidy and T21, 2) euploidy and T18, 3) euploidy and Turner and 4) T21 and OCA

system due to very low performance and combine the results of the three systems in a logical way ("euploidy & T21", "euploidy & T18", and "euploidy & Turner syndrome").

System 2 was therefore decided to involve two stages for distinguishing between euploidy, T21, and aneuploidy pregnancies. In Stage 1, each case is assessed independently by the three models explained above and classified as euploidy or aneuploidy (T21, T18 or Turner syndrome). This is done with the logical statement "If a case is classified as euploidy by all three models then this case is given a final classi-

fication as euploidy, otherwise is classified as OCA and is sent to Stage 2 for further examination". In Stage 2, all aneuploidies are reassessed by a 9-input model of "T21 & OCA" with a binary output (0 for T21 and 1 for OCA) and reclassified into T21 or OCA as shown in Fig. 2.7.

2.3 Results

We present our results in terms of detection rate, the accuracy, and Matthews correlation coefficient (MCC) (Matthews 1975). A correct classification of an abnormal case is called true positive prediction (TP), while a false classification of an abnormal case is called false negative prediction (FN). True negative (TN) and false positive (FP) are the correct and false classification respectively, for a normal case. The detection rate is defined as the correctly classified instances divided by the total population of each class. The accuracy is defined as the sum of true positives and true negatives divided by the total population. The MCC is a balanced measure of the quality of binary classifications in the range -1 and 1 and it is commonly used to describe the results of highly imbalanced class populations. A value of -1 represent a complete error of classification while a value of 1 represents perfect classification. A value of 0 shows random classification. It was introduced by the biochemist Brian W. Matthews in 1975 and it is defined as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.2}$$

2.3.1 System 1: Classification into two classes: euploidy and T21

The results of System 1 for fold 1 and for ANN, SVM (kernel 1 and 2) and k-NN are summarized in Table 2.6. The results of the three-fold cross validation for ANN are shown in Table 2.7. In table 2.8 we present the histogram of the output of the 9-input model where in the range 0 to 0.5 the system correctly identified as abnormal all 129 cases of T21 with a FPR of 3.9%. In the case of T21 pregnancies the average output value was 0.029 (standard deviation 0.074) and in 123 (95.4%) of cases the output was in the range of 0 to 0.2. In the euploidy pregnancies the average output value was 0.95 (standard deviation 0.15) and in 15,604 (92.3%) of cases the output was in the range of $>$0.8 to 1. The overall diagnostic yield of the system was true negative rate (TNR) of 96.1%, FNR of 0%, true positive rate of 100% and FPR of 3.9% (Fig. 2.6). Table 2.8 can become a practical tool because it can show for example that the FPR can be reduced considerably to 1% by simply reducing the threshold to 0.2, if the society is ready to support 4.6% of the undetected (FN) T21. This can also be interpreted as roughly 50 cases of T21 in 1000000 births.

Table 2.6: Accuracy and the Matthews correlation coefficient (MCC) of the ANN, SVM and k-NN of System 1 for the first validation dataset. System 1 was trained with euploidy and T21 cases. It was validated for the entire database including euploidy, T21 and OCA.

System 1	Accuracy	MCC
ANN	0.96	0.40
SVM 1	0.93	0.31
SVM 2	0.92	0.28
k-NN	0.92	0.28

Table 2.7: Detection rates of ANN of System 1 for the three validation datasets. System 1 was trained with euploidy and T21 cases. It was validated for the entire database including euploidy, T21 and OCA.

System 1	Euploidy	T21	OCA
Fold 1	96.1%	100.0%	27.6%
Fold 2	97.1%	93.9%	57.9%
Fold 3	97.2%	90.1%	65.8%

2.3.2 System 2: Classification into three classes: euploidy, T21 and OCA

System 2 is consisted of two subsystems (Stage 1 and Stage 2). In Stage 1 the system classifies an unknown case as euploidy or aneuploidy. In Stage 2, the aneuploidies are being further classified as T21 or OCA. (Fig. 2.7).

Classification into euploidy or aneuploidy by three models.

The network combined the 9-input model "euploidy & T21" (used in system 1) and the 6-input models "euploidy & T18" and "euploidy & Turner" syndrome. This network correctly classified as euploidy (concordance in all three systems) 15820 (93.6%) of the 16898 euploidy cases and as aneuploidy (in any one of the three systems) 196 (95.6%) of the 205 aneuploidy cases, including all 129 cases of T21 (100%), 34 (87.2%) of the 39 cases of T18, 11 (78.6%) of the 14 cases of T13, 8 (80%) of the 10 cases of triploidy and 11 (84.6%) of the 13 cases of Turner syndrome (Fig. 2.7).

Therefore, at the end of Stage 1, 1274 cases were classified as aneuploidy, including 1078 euploidy pregnancies (FPR 6.4%) and 15829 cases were classified as

Table 2.8: Performance of System 1, the 9-input model "euploidy & aneuploidy" where each case was quantified to a class between 0 and 1.

Output	Euploidy n=16,898	T21 n=129	OCA n=76
0 to 0.1	104 (0.6%)	115 (89.2%)	12 (15.8%)
> 0.1 to 0.2	80 (0.5%)	8 (6.2%)	3 (3.9%)
> 0.2 to 0.3	129 (0.8%)	3 (2.3%)	1 (1.3%)
> 0.3 to 0.4	138 (0.8%)	2 (1.6%)	2 (2.6%)
>0.4 to 0.5	203 (1.2%)	1 (0.8%)	3 (3.9%)
>0.5 to 0.6	192 (1.1%)	-	2 (23.6%)
>0.6 to 0.7	217 (1.3%)	-	2 (23.6%)
>0.7 to 0.8	231 (1.4%)	-	5 (6.6%)
>0.8 to 0.9	123 (0.7%)	-	7 (9.2%)
>0.9 to 1.0	15481 (91.6%)	-	39 (51.3%)

euploidy, including 15820 (99.9%) which were truly euploidy and 9 of the OCA. The detection rate of the euploidy of SVMs with kernel 1 and 2 in Stage 1 is 93.4% and 92.2% and for the aneuploidies 73.2% and 72.68% respectively. The k-NN classified correctly 91.7% of the euploidy and 74.2% of the aneuploidies.

The TPR and TNR for the networks trained with "euploidy & T21", "euploidy & T18" and "euploidy & Turner" are shown in Figs 2.8, 2.9 and 2.10 for the three evaluation datasets, labeled as fold 1, fold 2 and fold 3. We present the values of the TPR and TNR for different values of the threshold that were used to quantize the outputs of the networks and classify instances into the desired classes.

We tested several values of the threshold in the range between 0 and 1 with a step of 0.1. The first plot in Fig. 2.8 shows the results of the network "euploidy & T21". The TPR for the T21 aneuploidies is plotted with black continuous line and with black dashed line the TNR for the first evaluation dataset (fold 1). This network was used individually to construct system one, while it is also used as a subpart in system two.

The TPR reaches maximum rate 100.0% at a 0.48 threshold. The FPR at the maximum TPR is 3.9%. The results of the second and the third evaluation datasets are shown in the next two plots respectively in Fig. 2.8.

The results of the network "eupoid & T18" for the three validation datasets are shown in Fig. 2.9. The TPR of the T18 cases reaches maximum rate 89.8% at a FPR of 2.6% and threshold 0.98 for the first evaluation dataset. The TPR of the second

and third evaluation dataset at around 3% FPR is 92.5% and 90.2% respectively.

The results of the network "euploidy & Turner" for the three validation datasets are shown in Fig. 2.10. The TPR for the first evaluation dataset of the Turner cases reaches maximum rate 76.9% at a FPR of 1.3% and threshold 0.98. The TPR of the second evaluation dataset reaches maximum rate 100.0% at a FPR of 1.7%. The TPR of the third evaluation dataset reaches maximum rate 66.7% at a FPR of 3.0%.

Classification into T21 or OCA.

The 1274 cases classified as aneuploidy in Stage 1 were examined by the 9-input model "T21 & OCA" in Stage 2 (Fig. 2.7). The output values for all cases of T21 were 0 to 0.1, whereas the values of the OCA were mostly distributed near to 1. In 64 (95.5 %) of the 67 OCA the output was more than 0.1 and therefore three of these cases were wrongly classified as T21. The SVMs with kernel 1 and kernel 2 classified 98.5% and 97.7% of the T21 and 6.6% of the OCA k-NN classified 62.0% of the T21 and 32.9% of the OCA.

The accuracy and MCC of the ANN, SVM and k-NN for the System 2 in Stage 2 for the first fold validation are summarized in Table 2.9. The detection rates of the first, second and the third validation sets for the ANN are summarized in Table 2.10. It is shown from Table 2.9 that only ANN were able to separate T21 from the OCA while SVM and k-NN failed.

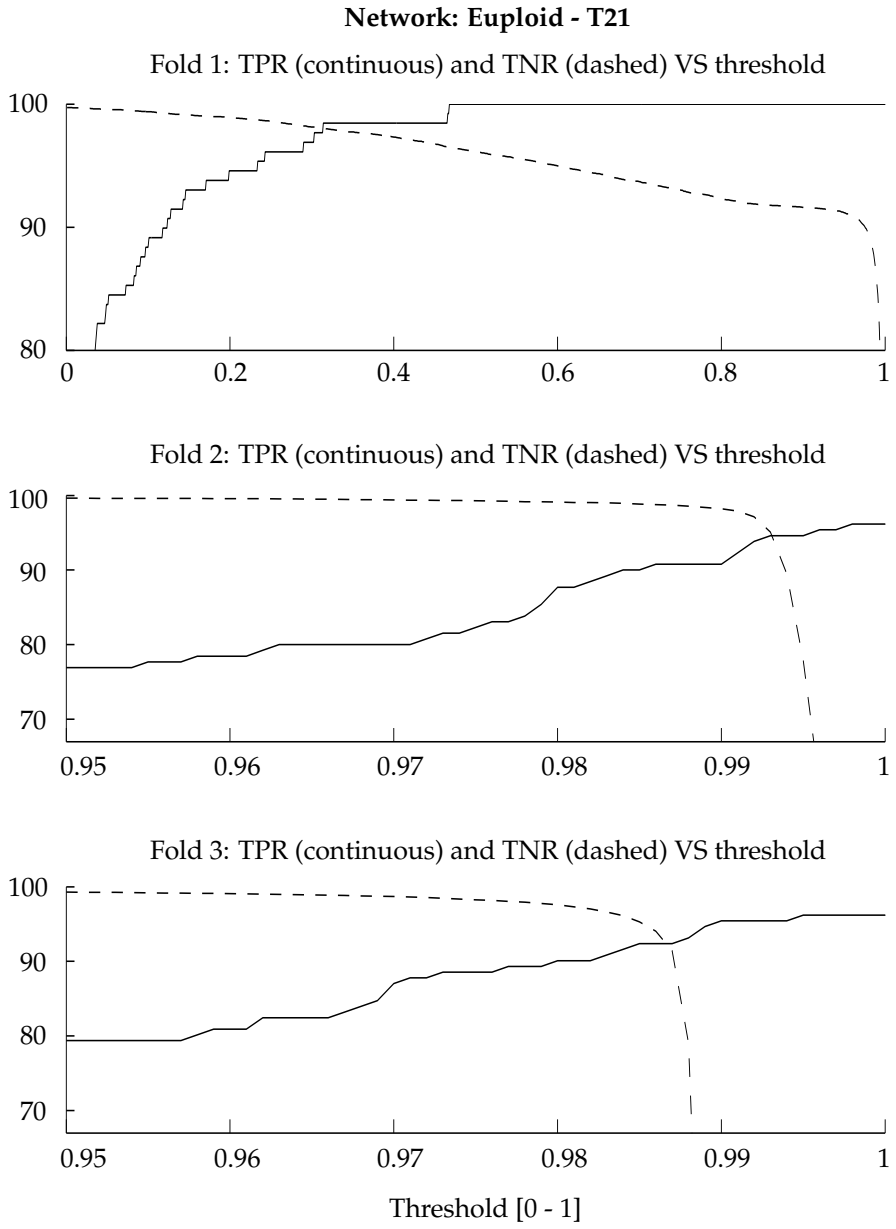

Figure 2.8: TPR (filled line) and TNR (dashed line) for the network "euploidy & T21". The three plots show the results for the three validation sets (fold 1, 2 and 3).

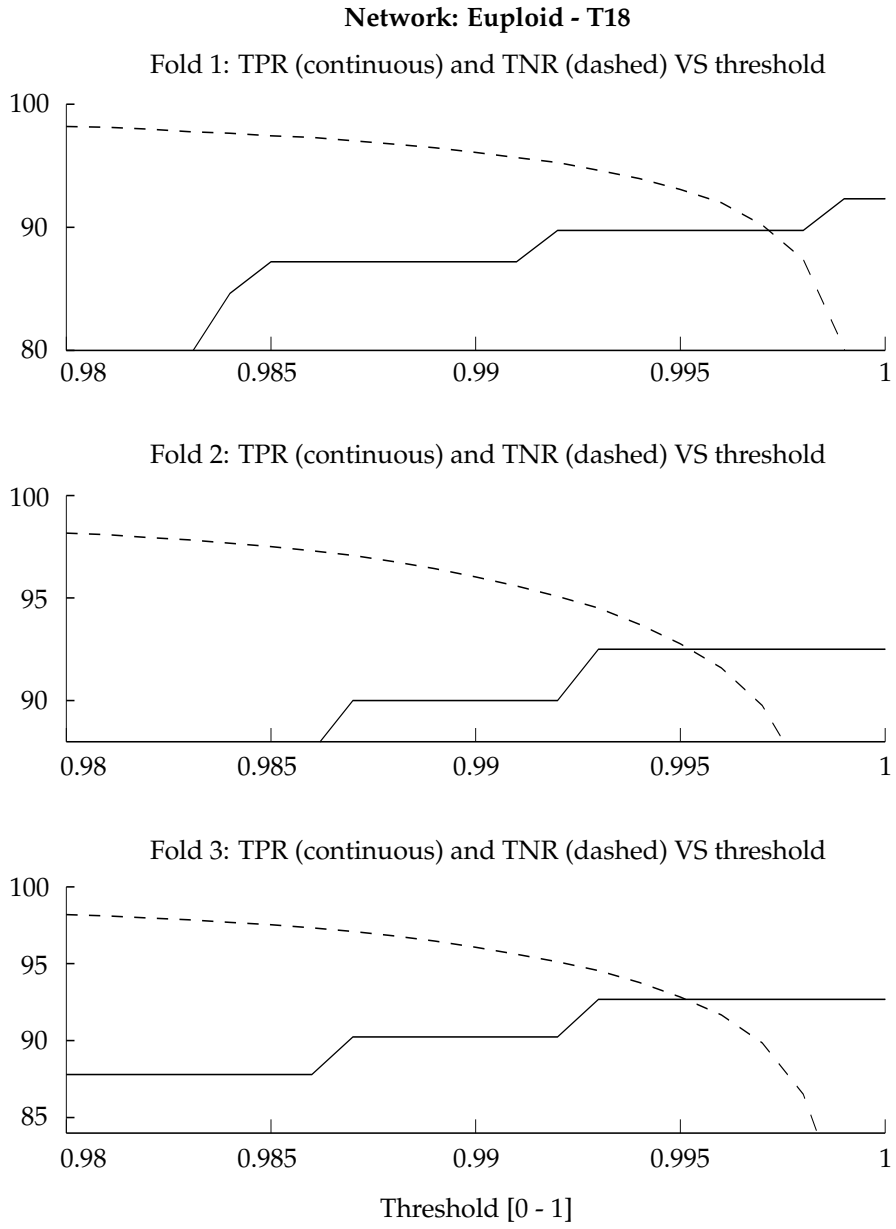

Figure 2.9: TPR (filled line) and TNR (dashed line) for the network "euploidy & T18". The three plots show the results for the three validation sets (fold 1, 2 and 3).

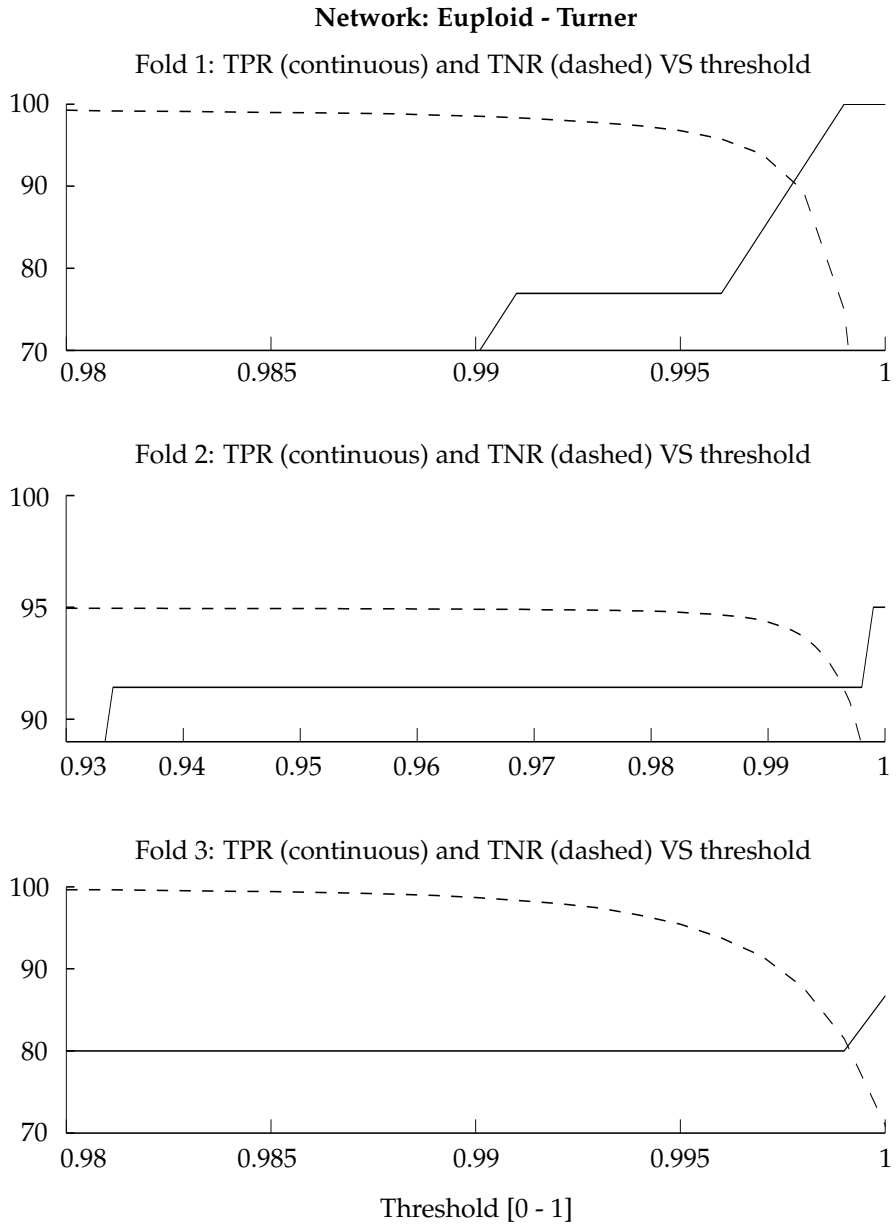

Figure 2.10: TPR (filled line) and TNR (dashed line) for the network "euploidy & Turner". The three plots show the results for the three validation sets (fold 1, 2 and 3).

Table 2.9: Accuracy and the Matthews correlation coefficient of the ANN, SVM and k-NN of System 2, Stage 2 for the first validation dataset. System 2 in Stage 2 was trained with T21 and OCA. It was validated for the entire database including euploidy, T21 and OCA.

System 2	Accuracy	MCC
ANN	0.94	0.88
SVM 1	0.64	0.13
SVM 2	0.64	0.11
k-NN	0.51	-0.05

Table 2.10: Detection rates of the ANN of the system 2 in phase 2. System 2 is consisted by one neural network build with T21 and OCA.

System 2 phase 2	Euploidy	T21	OCA
Fold 1	93.7%	100.0%	84.3%
Fold 2	96.7%	90.0%	57.9%
Fold 3	96.2 %	87.8%	44.3%

2.4 Discussion

The findings of our studies demonstrate the value of ANN schemes in the prediction of T21 and OCA from ultrasonographic and biochemical markers at 11-13 weeks of gestation.

A multitude of ANN structures, training procedures and evaluation strategies have been tried. In this study we used multilayer feed forward neural systems because these are proved to be the most suitable from the point of view of satisfactory generalization and diagnostic yield for such predictive systems. This was confirmed empirically by the authors after running several ANN models with different structures and parameters and observing their performance. The various multilayer networks of neurons were build and adjusted according to a set of parameters for each case of either euploidy or aneuploidy fetus in order to maximize the correct identification of each group. We have carried out a comparative study by using other classification techniques such as the SVMs and the k-NNs. The higher accuracy on the classification of euploidy, T21 and OCA was achieved with neural networks structures.

In the traditional approach to first trimester screening for T21 the a priori risk

is multiplied with the likelihood ratio of each sonographic and biochemical marker. Algorithms combining six parameters, including maternal age, previous history of aneuploidy, delta NT, serum free β-hCG MoM and PAPP-A MoM have been successfully applied to screening for T21 achieving a detection rate of about 90.0%, at a FPR of 5.0% (Nicolaides 2011). In specialist fetal medicine centres the performance of screening can be improved further, with an increase in detection rate to about 95.0% and a decrease in FPR to less than 3.0%, by the inclusion the additional sonographic markers of presence or absence of the fetal nasal bone and normal or abnormal blood flow across the tricuspid valve and in the ductus venosus (Nicolaides 2011).

In this study we initially attempted to develop a supervised ANN systems with six outputs: one each for the euploidy pregnancies and the five chromosomal abnormalities were included in the system. However, this was unsuccessful because the cases of T21 were separable but when the OCA were in the system, this separation was destructed and most of the aneuploidies were classified as euploidy.

Subsequently, we concentrated our efforts at developing neural network models with the intention of separating euploidy from T21 pregnancies. One model utilized six neurons representing the basic parameters of maternal age, previous history and the other 9 neurons representing the parameters shown in Table 2.5 columns 1 and 2 respectively. A large database consisting of 33619 euploidy and 279 T21 pregnancies was used to train the systems which were then used for testing the totally unknown database which included 16898 euploidy, 129 T21 and 76 OCA. pregnancies. The 6-input model correctly identified 92.3% of the cases of T21 at a FPR of 3.9% and the 9-input model detected all cases of T21 at a FPR of 3.9%.

The 6-input and 9-input "euploidy & T21" models recognized that 63.2% and 27.6% respectively, of the aneuploidies other than T21 were different from euploidy pregnancies. Subsequently, various attempts were made to improve the detection rate of the OCA by building neural networks that were trained to separate euploidy from each type of aneuploidy. The performance of neural networks attempting to separate T13 and triploid from euploidy pregnancies was low and these models were abandoned. A two-stage approach involving four neural networks was then used to achieve the best overall performance. In Stage 1, each case was assessed independently by three models ("euploidy & T21", "euploidy & T18" and "euploidy & Turner syndrome"). This two-stage approach correctly identified all cases of T21 and 84.3% of the OCA but at an overall FPR of 6.4%.

Like every methodology, ANN and the computational intelligence approach have relative advantages and disadvantages. Some of the significant advantages, when compared to the above methodologies the proposed approach are:

1. Every case is seen by the system as a string of parameter values and are thus processed and assessed simultaneously. At the same time data related to the fetus are seen and assessed together with data that is collected from the mother.

2. Each new case that has been observed during pregnancy can become a new

definite case once child delivery takes place. This new case can add to the acumen of the existing knowledge base by simply running the learning algorithm once the new case enters the database.

3. In addition to the above, advantages such as fault tolerance, generalisation handling, missing data handling, learning and inference mechanisms, are advantages inherited from computational intelligence.

At present most medical centres providing first trimester screening for T21 measure fetal NT and CRL and maternal serum free β-hCG and PAPP-A. In such centres the use of the proposed combined 6-input system, could correctly identify as aneuploidy about 93.0% of the cases of T21 and 63.0% of those with OCA, at a FPR of 3.9%. This performance of screening compares favourably with the 90.0% detection rate of T21, at a FPR of 5.0%, achieved by the traditional algorithms for screening (Nicolaides 2011). Nevertheless, in all cases classified by the neural network as being suspicious of T21 invasive testing by chorionic villous sampling or amniocentesis would still be necessary to distinguish between the euploidy and aneuploidy pregnancies and in the diagnosis of the exact type of aneuploidy.

In fetal medicine centres, with expertise in assessing the fetal nasal bone and Doppler flow across the tricuspid valve and in the ductus venosus in addition to the measurements of fetal NT and CRL and maternal serum free β-hCG and PAPP-A, there are two options on the use of ANNs. The first is to use a 9-input "euploidy & T21" model which could correctly identify as aneuploidy all cases of T21 and 27.6% of those with other major aneuploidies, at a FPR of 3.9%. Alternatively, a two-stage approach involving four neural networks can be used which could also correctly identify as aneuploidy all cases of T21 and 84.3% of those with OCA, but at an increased FPR of 6.4%. Since in all cases classified by the neural networks as being suspicious of T21 invasive testing would be necessary to distinguish between the euploidy and aneuploidy pregnancies it is likely that the first option, with a substantially lower FPR, would be preferred by the parents and would also be more cost-effective. Although the second option identifies more of the OCA, unlike T21 these conditions are highly lethal either in utero or in the neonatal period and they are associated with abnormalities that can be easily detected by ultrasonography. These include holoprosencephaly, exomphalos and megacystis in trisomies 18 and 13 (Kagan et al. 2010), large cystic hygromas in Turner syndrome (Kagan, Wright, Spencer, Molina and Nicolaides 2008) and either an enlarged partially molar placenta or small placenta but severely growth restricted fetus with pronounced wasting of the body and sparing of the head in triploidy. Since the prevalence of these defects is less than 0.1% the effect on the overall proportion of pregnancies requiring an invasive test would be minimal.

2.5 Conclusion

We have presented a non-invasive prenatal diagnosis of chromosomal abnormalities in the first trimester of the pregnancy. The collection of the database, took several years and it covers a wide area of population such as age range, ethnicity, information whether the mother is alcoholic, drug addicted, cases with previous history. The population of the cases in our database, ensures statistical confidence of our results, compare to databases used in similar work of other groups. We believe that medical diagnosis systems that classify instances based on statistical methods should provide representative databases with a convincing number of population. Also, the results should be presented with cross validation, ensuring their robustness. The most important however is the fact that our system identifies and predicts OCA than T21, such as T18, T13, Turner syndrome and Triploidy. The diagnosis is done in the presence of the pregnant woman with a computer system in the doctor's office by basically using routine examination data within a negligible time and with low financial cost.

We achieved with ANN a 100.0% detection rate of T21 with FPR of 3.9% and 84.3% of the OCA with a FPR of 6.4%. These results do not yield percent classification neither 100% accuracy since there is still a FPR of 6.4%. Even though the main purpose of this work was not to compare the various computational intelligence paradigms, we have repeated the same experiments with SVM and k-NN and it was experimentally concluded that the best results for this problem would be achieved with the ANN. More precisely, SVM and k-NN yield similar results with lower accuracy than the proposed ANN structure of both systems 1 and 2. It is also prominent that both SVM and k-NN were not able to distinguish the T21 from the OCA.

For future work we need to emphasise our investigations on exploring other neural network schemes such as recurrent networks and the possibility of using parameters from the father. Preliminary results show that it is worth investigating the substitution of the parameter values in MoMs with the actual raw values. It is also worth mentioning that the maternal cell-free DNA screening method does not use any information from the mother or the father; such as the age. This important information should be included in their methodology and the analysis of their results.

Accepted for publication as:

A. Neocleous, K. Nicolaides, and C. Schizas, "Intelligent Non-invasive Diagnosis of Aneuploidy: Raw Values and Highly Imbalanced Dataset," *IEEE Journal of Biomedical and Health Informatics*, 2016.

Chapter 3

Non-invasive Diagnosis of Aneuploidy: Raw Values and Highly Imbalanced Dataset

Abstract

The objective of this work is to introduce a non-invasive diagnosis procedure for aneuploidy and minimize the social and financial cost of prenatal diagnosis tests that are performed for fetal aneuploidies in an early stage of pregnancy. We propose a method using artificial neural networks trained with data from singleton pregnancy cases, while undergoing first trimester screening. Three different datasets[1] with a total of 122362 euploid and 967 aneuploid cases were used in this study. The data for each case contained markers collected from the mother and the fetus. This study, unlike previous studies published by the authors for a similar problem differs in three basic principles being, a) the training of the artificial neural networks is done using the markers' values in their raw form (unprocessed), b) a balanced training dataset is created and used by selecting only a representative number of euploids for the training phase, and c) emphasis is given to the financials and suggest hierarchy and necessity of the available tests. The proposed artificial neural networks models were optimized in the sense of reaching a minimum false positive rate and at the same time securing a 100% detection rate for Trisomy 21. These systems correctly identify other aneuploidies (Trisomies 13&18, Turner, and Triploid syndromes) at a detection rate >80%. In conclusion, we demonstrate that artificial neural network systems can contribute in providing non-invasive, effective early screening for fetal aneuploidies with results that compare favorably to other existing methods.

3.1 Introduction

The early diagnosis of fetal aneuploidies in the first trimester of pregnancy can be achieved with amniocentesis or Chorionic Villus Sampling (CVS). However, such methods are invasive and they carry a risk of infections, fetal damage during the examination and miscarriage rates of about 0.4% for amniocentesis and 1.1% for CVS (Enzensberger et al. 2012).

[1]The dataset can become available for academic purposes by communicating directly with the authors.

In Europe, the cost of an amniocentesis test varies between 300 and 1000 euros. The cost of the loss of a life due to amniocentesis is tremendous. Therefore, it is important to reduce the false positive rate (FPR) as much as possible, but at the same time be able to detect all, or most of the aneuploidies. Previously, the overall cost for the detection was much higher because pregnant women were advised to go for amniocentesis based only on their age e.g > 35; which not only increased the FPR beyond 25% and the risk of unexpected aneuploidy births much greater than zero since in many cases the cost for the invasive test could not be afforded and/or the risk for miscarriage could not be taken by the parents.

As alternative, non-invasive methods have been proposed by Snijders et. al (Snijders et al. 1999, Snijders et al. 1998), Kagan et.al. (Kagan, Wright, Baker, Sahota and Nicolaides 2008, Kagan, Wright, Spencer, Molina and Nicolaides 2008) and Spencer et. al. (Spencer et al. 1999, Spencer et al. 2003). Essentially, a risk for aneuploidy is estimated based on a prenatal examination test that is performed to every pregnant woman in the first trimester of pregnancy. In the literature, the most relevant markers from the prenatal examination are the following: maternal age, serum free β-hCG, pregnancy-associated plasma protein-A, nuchal translucency thickness, nasal bone, tricuspid flow, and ductus venosus flow. A Statistical Mixture Model (SMM) is used in(Snijders et al. 1999, Snijders et al. 1998, Kagan, Wright, Baker, Sahota and Nicolaides 2008, Kagan, Wright, Spencer, Molina and Nicolaides 2008, Spencer et al. 1999, Spencer et al. 2003, Nicolaides 2005, Maiz et al. 2009, Huggon et al. 2003) as an estimator for the risk of trisomy 21 (T21). Outcomes that are rated as "high risk" are suggested to follow invasive test. Currently, the detection rate for the T21 of the abovementioned methods is 95% at a 5% FPR.

In the last years, another non-invasive method has gained particular attention in the scientific community. A sample from the maternal blood is used to isolate the plasma via double centrifugation. Then, the circulating cell-free DNA is sequenced from the maternal plasma, using state-of-the-art equipment. The density of the DNA sequences are normalized and distributed for every chromosome separately for euploid, T21 and T18. A standard statistical classification technique such as Z-transform or t-test is applied to the euploidy and aneuploidy distributions of the chromosomes 21 and 18 to estimate the probability for aneuploidy of an unknown case (Ashoor et al. 2012, Papageorgiou et al. 2011, Zhang et al. 2015, Sahoo et al. 2006, Kagan et al. 2015).

In this work we propose the use of a computational intelligence approach using artificial neural networks (ANNs) that have the ability to digitally store the information from a training dataset, in a similar manner to the biological function of the human brain (Levine 2000). In Section 1 we have showed that ANNs can achieve 0% false negative rate (FNR) for T21 at 3.6% FPR.

3.1.1 Statistical Mixture Model

A prenatal examination is performed to the pregnant woman between the 9th + 3 and 11th week + 6 days of gestation. From the maternal blood, the concentrations of the two biochemical markers a) pregnancy-associated plasma protein-A (PAPP-A) and b) the serum free β -hCG are measured. Fetal ultrasonographic markers include the levels of appearance of a subcutaneous collection of fluid behind the fetal neck, called Nuchal Translucency (NT). The amount of the NT is statistically increased in fetuses with trisomy 21 (Kagan, Wright, Baker, Sahota and Nicolaides 2008, Kagan, Wright, Spencer, Molina and Nicolaides 2008, Spencer et al. 1999). Other sonographic markers are measured such as the Crown Rump Length (CRL) and the Nasal Bone (NB). The CRL is a physical measurement of the length in millimeters between the neck and the bottom of the buttocks of the fetus. The NB indicates the presence or the absence of the fetal NB and the obstetrician marks it at the time of the examination as normal or abnormal respectively.

It has been proposed in (Kagan, Wright, Baker, Sahota and Nicolaides 2008, Kagan, Wright, Spencer, Molina and Nicolaides 2008, Spencer et al. 1999, Maiz et al. 2009, Huggon et al. 2003) that the biochemical markers β-hCG and PAPP-A and the NT increase their separability strength between euploid and aneuploid. Additionally, the values of the biochemical markers are normalized with their multiple of their medians (MoM).

The MoMs is a data normalization method that has been found effective in medical data (Kagan, Wright, Baker, Sahota and Nicolaides 2008, Kagan, Wright, Spencer, Molina and Nicolaides 2008) and is a measure of how many times an individual test result deviates from the median. The term "multiples" refers to this measure. The MoMs are calculated as follow: first the data are clustered in three categories based on the gestational age at the time of the examination: a) 9-10, b) 10-11 and c) 11-12 weeks of gestation. Then, for every category the median value of the biochemical markers for the euploid cases is calculated. Finally, the raw observation is divided by the respective mean value of the specific gestation age. The NT did not respond positively to the MoM values and it is transformed into the delta NT that is a measure of the deviation from its euploid zero median.

The classification is done as follow: for every marker, a risk is estimated based on the Gaussian distribution of the respective marker values and it is multiplied with the maternal age related risk to yield a final result. The Ductus Venosus Flow (DV) and the Tricuspid Valve Flow (TF) may be used to increase the detection rate and reduce the FPR (Maiz et al. 2009, Huggon et al. 2003).

3.1.2 Cell-Free Fetal DNA Test

In the literature, the most promising results within the non-invasive methods are achieved with the cell-free fetal DNA test. In (Ashoor et al. 2012), it is reported that a perfect isolation of 300 euploid and 50 cases of T21 is achieved, together with a 98% detection rate for the trisomy 18 (T18). This method returned no result in three euploid cases (about 1% of the population). Another recent study (Papageorgiou et al. 2011) reports 100% detection rate of 14 cases of T21 at 0% FPR with 26 euploid. The results in (Zhang et al. 2015) show that significantly lower FPR can be achieved compared to the SMM method that is currently used. Particularly 146958 cases have been studied including 726 T21, 170 T18 and 22 trisomy 13 (T13) for which outcome data were available in 112,669 (76%). The overall Sensitivity and Specificity are 99.02% and 99.86% respectively. Other studies (Sahoo et al. 2006, Kagan et al. 2015) report similar results.

Even though the results of the cell-free fetal DNA approach are significantly better than other non-invasive methods, we identify some drawbacks. First, the test does not yield immediate results. It requires specialized doctors, expensive equipment and laboratories. Therefore, it is a method that cannot be applied in one visit. Another drawback of this method is that the DNA sequence is statistically visualized and the classification is done using a threshold on the probability of aneuploidy risk. This means that the results need to be cross-validated in several standard ways e.g. 3-fold or leave-one-out cross-validation to make sure that are consistent. Finally, the identification of other chromosomal abnormalities (OCA) such as Turner syndrome or triploidy are not identified. However, a combination between the existing non-invasive methods may yield to an optimized solution to the problem. A combination of two non-invasive methods, the prenatal screening and the maternal blood cell-free fetal DNA testing is proposed in (Kagan et al. 2015).

3.1.3 Machine Learning

ANNs have been widely used in medical applications for the prediction of cancer(Schnorrenberg et al. 1997, Schizas and Pattichis 1997, Neves et al. 2015), Parkinson disease (Er et al. 2016) and other serious diseases (Wang et al. 2015, Yoo et al. 2016, Liu et al. 2015). The major difference between statistical methods and ANNs in classification is that ANNs have the ability to learn and store information by examples that are presented one by one. In other words, statistical information such as the distribution, the mean and standard deviation values are not significant. This is done in a similar process as the biological neural networks store and process information in the brain.

The research question of study is to examine the potential value of ANNs in the prediction of the risk for T21 and OCA from ultrasonographic and biochemical

markers at the early stage of gestation. Additionally, we make experiments to test the possible contribution of normalized data over the raw data. Furthermore, we question if the use of balanced training sets between euploid and aneuploid achieve more reliable and consistent results. Finally, we explore the contribution of different combinations of input markers. The objective of this study is to build a system that will identify 100% of the T21 at the lowest FPR possible, using the most reliable type of markers for the training of the ANNs, being raw or normalized, balanced or imbalanced training sets and the combination of the input markers.

In Section 5.2 we present our methodology which includes a description and analysis of the available data, the ANN structures, the cross-validation approach for testing the models, the normalization of the data and the procedure for creating balanced training sets. In Section 3.3 we present our results and we conclude with Sections 3.4, 3.5 and 3.6.

3.2 Methods

Three different datasets are provided by the Fetal Medicine Foundation (FMF) and used in this study. The first (*Dataset A*), consists of 51001 cases of pregnant women that followed a prenatal examination within the first trimester of pregnancy, and similarly the second (*Dataset B*) and the third (*Dataset C*) of 29999 and 42329 cases respectively. All the samples were very carefully collected and thoroughly tested, and they are maybe the largest complete dataset in existence today for this kind of study.

We used part of *Dataset A* as training set for performing a grid search of ANNs over a set of parameters such as the hidden units, activation functions and training epochs, as shown in Fig. 3.1. Then, we selected the ANN that yielded the best results on the remaining cases that formed the test set of *Dataset A*. All of our experiments were then done using *Datasets B* and *C* that we call "testing datasets". The ANNs were built using the parameters of the ANN that was identified as the optimal in the grid search.

We proceeded with our research approach by implementing three different experiments. The first experiment was done for identifying the optimal combination of markers that are needed as input to our system. We created eight groups of markers and we built an ANN for each one of them to compare their performance. The second experiment was done for testing whether the use of normalized data values for the markers outperform the use of raw values. The third experiment dealt with a more technical question concerned with the imbalanced nature of the datasets, which is due to the very low percentage of aneuploidies in the datasets. A data reduction technique was carried out for reducing the population of euploid cases during the training phase. Several representative ANNs were developed and tested, both with balanced and imbalanced datasets and the results were compared.

3.2.1 Data

The populations of euploid and aneuploid of each dataset are shown in Table 3.1. The vast majority of the cases are euploid creating a highly imbalanced situation between the euploid and aneuploid cases.

The number of markers/features that were available for our dataset is 22. Most of them are related to the physiological and historical data of the pregnant woman, such as the history of aneuploidy in previous pregnancies, smoking or drug habits, symptoms of hypertension, way of conception, ethnicity, etc. Other markers of apparently greater importance since they are taken during pregnancy are the biochemical PAPP-A and β-hCG and the ultra-sonographic markers, DV and the absence or presence of the fetal NB.

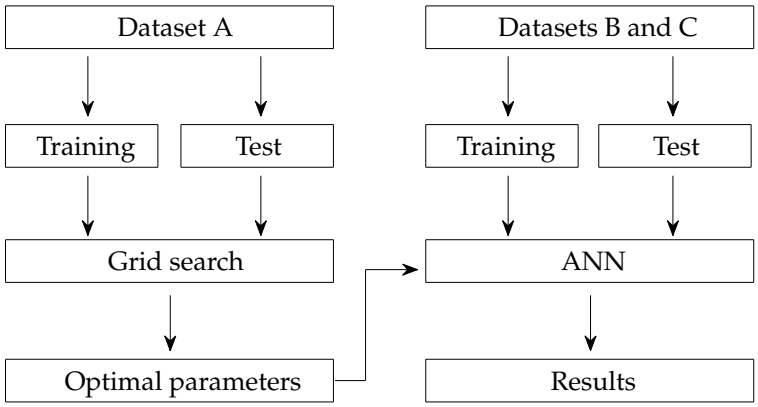

Figure 3.1: All of the three datasets are divided into training and test sets. The training set of the Dataset A is used to perform a grid search over different parameters of the ANN. We use the parameters of the network that yielded the best results on the test set of Dataset A to perform experiments using the Datasets B and C.

Table 3.1: Euploid and aneuploid populations of the three datasets used.

Dataset	Euploid	T21	T18	T13	Triploidy	Turner
A	50517	408	39	14	10	13
B	29790	124	42	10	14	19
C	42055	152	60	22	14	26

3.2.2 Artificial Neural Networks

The major advantage of the use of ANNs compared to other statistical approaches for classification tasks is their ability to learn by examples. A typical architecture of an ANN has one input layer, one or more hidden layers and one output layer. Each layer has a number of nodes that are connected to each other via a weight that represents the synapse in the biological neural networks. The first layer consists equal nodes to the number of input markers. The number of nodes in the hidden layers is a parameter and it has to be optimized manually, according to the problem under study. The last layer (output layer) contains one or more nodes, depending on the number of classes. In tasks with two classes, it is commonly used one node. The value of the output layer is finally passed through a step function where a cut-off point binarizes the output into 0 or 1.

In the training phase of an ANN, all the examples are presented to the nodes

Table 3.2: Cross validation. Training and validation sets for the three datasets used.

	Training		Validation		
Dataset	Euploid	T21	Euploid	T21	OCA
A	33619	279	16898	129	76
B	20782	87	9008	37	85
C	31225	100	10830	52	122

of the input layer one by one. The values of the input pattern are first multiplied. Then, all the products between the input values and the weights of every node are summed and passed to the nodes of the first hidden layer, through an activation function. The information from the hidden layer to the next hidden layers and the output layer is passed in a similar way as from the input layer to the hidden layer. One epoch is considered when all the examples are seen by the network and processed through the hidden layer(s) and the output layer. After every epoch, the weights of the ANN are updated based on a cost function that is calculated as an error function between the known target and the output value of the ANN. One common error function that is used in feed-forward backpropagation networks is the Mean Squared Error (MSE). The training of an ANN converges when the MSE is below a certain threshold, or when the specified epochs are reached. In our experiments we used feed-forward backpropagation networks with one hidden layer and one node in the output layer. The weights are initialized randomly and the learning rate was set at 0.3. We used 500 epochs for training the ANNs.

3.2.3 Cross Validation

The three datasets were split in two sets each of 70% and 30%. The first set is used for training the ANNs. The second set is kept away from the training procedure as the validation set. The number of cases in each training and validation sets for the three datasets used are shown in Table 3.2.

3.2.4 Marker Selection

After the consultation of the doctors that are involved in this research, we have made experiments with the aim of minimizing the required number of prenatal examinations that a pregnant woman is requested to perform. Two groups of markers namely "short" and "long" were examined as a potential two-stage screening for aneuploidy. The "short" group is consisted of the maternal age, the biochemical β-hCG and PAPP-A, and the fetal NT. In the "long" group we included three additional markers, the DV, the TF and the absence or presence of the NB that are

Table 3.3: Short and long groups of input markers that are used as inputs to the neural network. The abbreviations MA, CRL, NT, DV, TF and NB stand for maternal age, crown rump length, nuchal translucency, ductus venosus, tricuspid flow and nasal bone, respectively. The word YES indicates that the specific maker is used in the respective group. Similarly, the word NO indicates that a marker is not used.

	Long				Short			
Markers	9	8a	7	8b	6	5a	4	5b
MA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CRL	Yes	Yes	No	No	Yes	Yes	4	5b
PAPP-A	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
β-hCG	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prev. T21	Yes	No	No	Yes	Yes	No	No	Yes
NT	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DV	Yes	Yes	Yes	Yes	No	No	No	No
TF	Yes	Yes	Yes	Yes	No	No	No	No
NB	Yes	Yes	Yes	Yes	No	No	No	No

extracted from an additional special ultra sound examination.

From the two groups of markers, we created eight combinations that are used as input to the ANN. In Table 3.3 we present the combinations used in both "short" and "long" groups. In the second row we show the ID of every network that corresponds to the figures in Section 3.3.

3.2.5 Data Normalization

The conversion of the biochemical markers PAPP-A and β-hCG into their MoMs, and the NT into delta NT has been proposed by Kagan et. al (Kagan, Wright, Baker, Sahota and Nicolaides 2008) as a step in their methodology for the patient-specific risk for trisomy 21. The authors in (Kagan, Wright, Baker, Sahota and Nicolaides 2008) show that the values of the biochemical markers and the NT are correlated with the gestational age at the time of the examination.

In Fig. 3.2, we superimpose the raw (black dots) and the MoM (grey dots) values for the biochemical marker PAPP-A of all the cases of *Dataset A*. We note that the data were first sorted based on the gestational age. In X-axis, we plot the gestational age in days to better visualize the effect that the raw values of PAPP-A are correlated

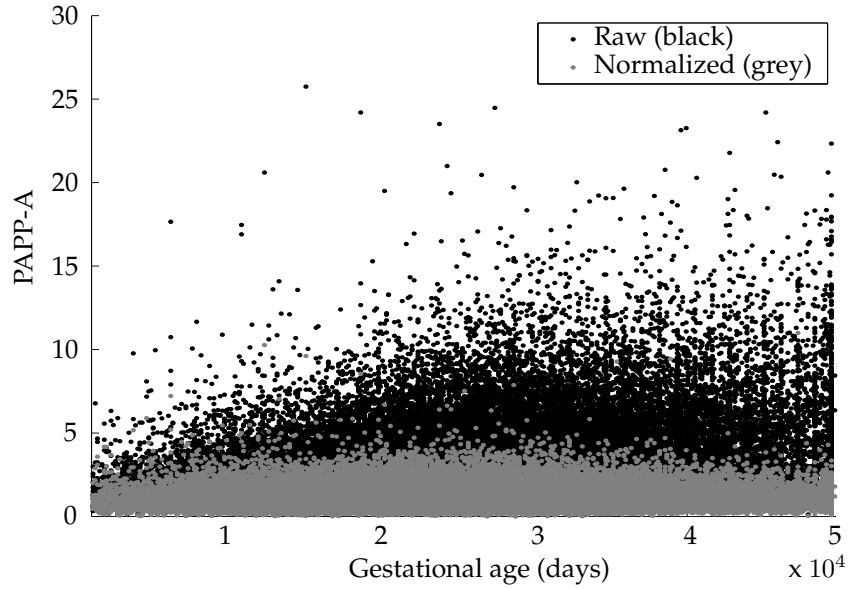

Figure 3.2: The Raw and normalized (MoM) values of the PAPP-A of all the cases that consist the *Dataset A*. The values are sorted in ascending order based on the gestational age. It is shown that the raw values of the PAPP-A increase with gestational age, while the normalized values are distributed around the average value of 2.5.

with gestational age. It is shown in Fig. 3.2 that the values after normalization are distributed in a lower range and the correlation to the gestational age is lost.

3.2.6 Data Reduction

The datasets used in this study are highly imbalanced: the euploid cases occupy more than 99% of the total population. In machine learning, the use of imbalanced populations for training may cause several technical problems. For instance, Bayesian classification uses an a-priori risk that is based on the population of each class. ANNs adjust their weights according to the MSE of each epoch, as explained in Section 3.3. Since the MSE is global for both minority and majority classes, the false predictions of the minority class are not influencing significantly the total MSE. In Fig. 3.3 we present the MSE of two neural networks trained with a) imbalanced (solid line) and b) balanced (dashed line) sets for 500 epochs. It is shown that the MSE of the balanced training set is significantly lower.

Several approaches for creating balanced training sets are proposed in the literature. One way is to choose representative instances from the major class to reduce

Table 3.4: Cluster map from the k-means algorithm.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Input	5084	15479	1619	227	11210
Reduced	76	231	25	4	67

the population for training. Another way is to artificially create data to increase the population of the minor class. Creating artificial cases for medical data is a difficult task due to the unpredictable correlation of the markers. For instance, the CRL and the NT are correlated with the gestational age at the time of the examination. This relationship is not fully understood and we are uncertain if we can model artificial data that will follow these specific pattern relations.

In this study, we choose the first approach to reduce the population of the major class. First, we apply unsupervised clustering to the euploid cases using the k-means algorithm with five prototypes. The number of the prototypes was selected by applying the Elbow method (Thorndike 1953). Then, for every cluster we identify the k-nearest neighbors of the respective prototype. The number of k is defined automatically and it is proportional to the length of the respective cluster with the length of the total euploid population:

$$k = target\_population \frac{Size(cluster_k)}{Size(euploid)} \tag{3.1}$$

In the second row of Table 3.4, we present the distribution of the k-means outputs that were built with the combination of the biomarkers and the ultrasonographic markers. In this example, we choose to reduce the euploid for training from 29790 to 503 cases. In the last row of Table 3.4 we present the number of the representative cases that are collected from every cluster, using the Eq. 3.1.

In Fig. 3.4 we present a 2D plot of the biochemical markers β-hCG (x-axis) and PAPP-A (y-axis) for the 5084 cases of the first cluster that are shown with black dots. The prototype of the cluster is shown with a white star and the 76-nearest neighbors to the prototype are shown with grey crosses. It is emphasized that in this representation, the cases that are chosen as representatives are not necessarily the ones that are closer to the prototype. This is due to the contribution of the other markers used in the training of the k-means. That would be the case if k-means were trained solely with the two biomarkers β-hCG and PAPP-A.

3.2.7 Evaluation Protocol

Choosing the best ANN architecture for a problem under study is not an easy task and it is usually done empirically by the system designer (chapter 2). In an attempt

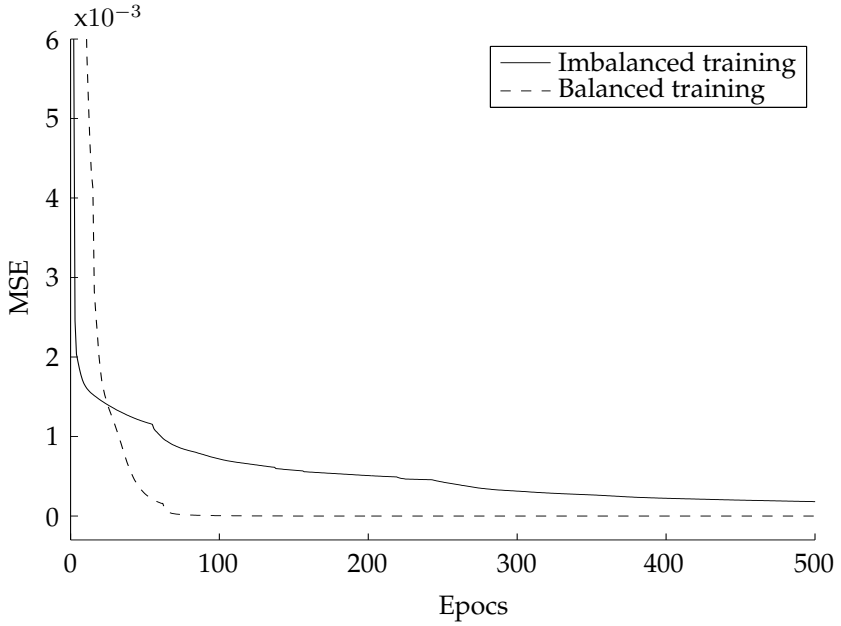

Figure 3.3: Mean Squared Error (MSE) of the neural network that was trained for 500 epochs. Dashed and solid lines show the MSE of balanced and imbalanced training sets.

to choose the optimal neural network architecture and parameters, we followed a grid search approach using the training set of *Dataset A* to construct 24 neural networks by changing the number of neurons in the hidden layer from 5 to 60, with step of 5 neurons, the logistic and the hyperbolic tangent transfer functions. The network that returned the best results on the validation set of *Dataset A* was used to construct the networks for the *Datasets B* and *C* that were used in all of our experiments. We use 50 nodes in the hidden layer with logistic activation function.

3.3 Results

The findings of our experiments are summarized in Figs. 3.5, 3.6 and 3.7 and Tables 3.5, 3.6 and 3.7. In Fig. 3.5 we show the results of the Dataset B and we present the FPRs at 0% FNR for T21 of the eight ANN models that were built using different combinations of the input markers. Additionally, we visualize the difference of the results between the networks trained with normalized and raw data with solid and dashed lines respectively. The difference between the performance of the raw and normalized ANNs is statistically significant for *Dataset B* but not for *Datasets A* and *C*. In *Dataset B*, as shown in Fig. 5, the "short" marker group (ID: 4, 5a, 5b and

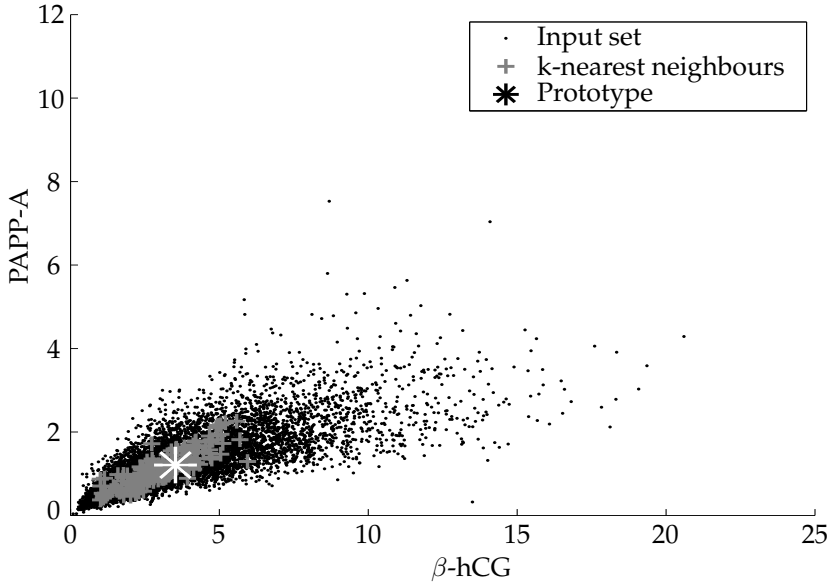

Figure 3.4: 2D plot of the biomarkers β-hCG and PAPP-A. The euploid cases of the first k-means cluster are presented with black dots. The prototype is shown with a white star and the 76-nearest to the prototype are shown with grey dots.

6) returned relatively high FPR (>20%) while the long group (ID: 7, 8a, 8b and 9) yielded significantly lower FPR.

In Fig. 3.6 we present the FPRs of the same models as in Fig. 3.5 at the 75% diagnostic rate of the OCA. The "short" group returned a FPR of 4.5% (in average of four ANNs) and of 8.5% for the models built with the normalized and raw values respectively. There is no significant difference between the FPRs (average of 0.2%) of the "long" normalized/raw marker groups.

In an attempt to visualize the difference of the performance between the balanced and imbalanced training sets, we calculate the sensitivity and the 1-Specificity for different cut-off values as explained in Section and we present the results of the network built with raw values and the marker set with seven inputs in a Receiver Operating Characteristic (ROC) curve (Fig. 3.7). The sensitivity and the Specificity are defined as shown in Eqs 3.2 and 3.3.

$$sensitivity = \frac{TP}{TP + FN} \tag{3.2}$$

$$specificity = \frac{TN}{TN + FP} \tag{3.3}$$

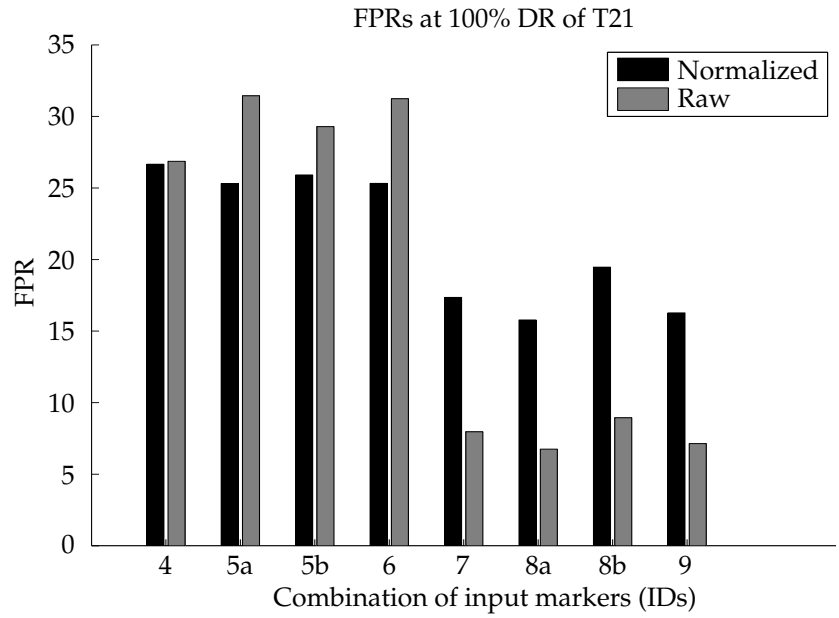

Figure 3.5: FPRs of the ANNs built with normalized (black bars) and raw (grey bars) values of *Dataset B* at a T21 DR of 100%. The results of the different input markers combinations are shown in the x-axis.

Where: TP, FN, TN and FP are abbreviations of the true positive, false negative, true negative and false positive respectively.

In Fig. 3.7 we superimpose the ROC curves for the networks built with raw marker values for imbalanced and balanced training sets using the "long" feature set with seven inputs. The results of the balanced and imbalanced training sets are distinguished with solid (black for T21) and dashed lines respectively (grey for OCA). It is shown that there is a significant difference between the DRs of the OCA built with balanced and imbalanced datasets. The DR of the T21 has no statistical difference.

In Tables 3.5, 3.6 and 3.7 we show the results of the networks built with the balanced and imbalanced training sets and the raw and the normalized seven input markers for the three datasets. In the first column of every table we present the type of the network (balanced or imbalanced training sets, and raw or normalized markers). In the second column we show the FPR and in the last two the DRs for the T21 and the OCA respectively.

The best results of the experiments done with *Datasets B* and *C* were achieved with the ANNs built with the 7 markers group and raw balanced data (8% and 4.8% FPR). The performance of the ANN built with the normalized *Dataset B* is signifi-

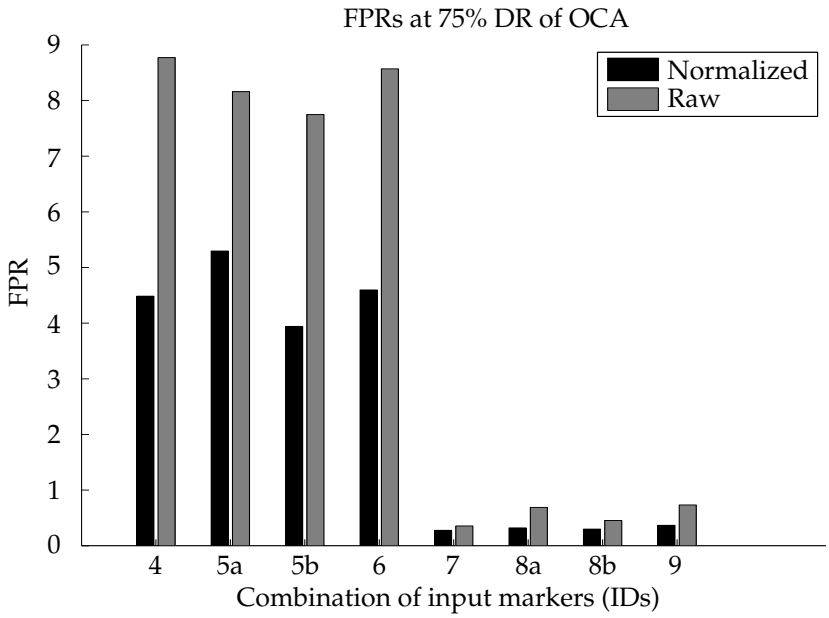

Figure 3.6: FPRs of the ANNs built with normalized (black bars) and raw (grey bars) values of *Dataset B* at a OCA DR of 75%. The results of the different input markers combinations are shown in the x-axis.

Table 3.5: FPR and DR for T21 and OCA for the networks built with *Dataset A*. The data are referred to the 7 markers group and raw data.

Dataset A	FPR	T21	OCA
Imbalanced - Normalized	2.7%	100.0%	7.9%
Imbalanced - Raw	7.9%	100.0%	32.9%
Balanced - Normalized	10.0%	100.0%	73.7%
Balanced - Raw	8.3%	100.0%	69.7%

cantly weaker with a difference more than 10% of FPR, compared to the ANNs built with the raw data. The same experiment was done with *Dataset C* with no difference on the results between raw and normalized data. With the results of *Dataset B*, we suggest that the normalization of the data could be avoided.

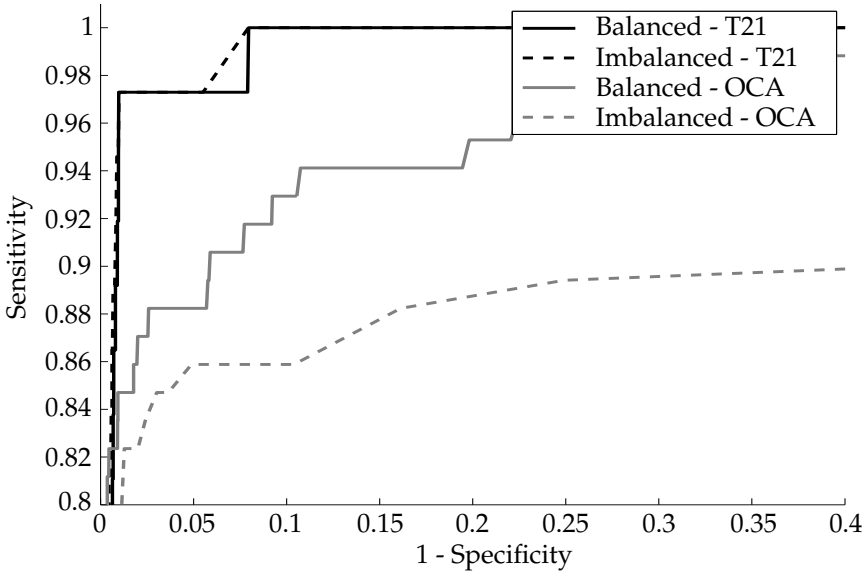

Figure 3.7: ROC curves of the models build with balanced (solid lines) and imbalanced (dashed lines) datasets for the euploid-T21 (black lines) and euploid-OCA (grey lines). The data are referred to the 7 markers group and raw data.

Table 3.6: FPR and DR for T21 and OCA for the networks built with *Dataset B*. The data are referred to the 7 markers group and raw data.

Dataset B	FPR	T21	OCA
Imbalanced - Normalized	34.2%	100.0%	90.6%
Imbalanced - Raw	7.9%	100.0%	83.5%
Balanced - Normalized	17.4%	100.0%	94.1%
Balanced - Raw	8.0%	100.0%	89.4%

3.4 Discussion

In this work we demonstrate the effectiveness of the ANNs schemes as a potential classifier for the diagnosis of the T21 and OCA, in the early stage of a pregnancy. Our principle aim is to build models that ensure false negative classifications of T21 at the lowest possible rate. The best results were achieved with the 7 markers group (*Dataset A*: 100% DR of T21 and FPR of 2.7%) with normalized imbalanced data. However, the *Dataset A* was used as a training set to perform a grid search over

Table 3.7: FPR and DR for T21 and OCA for the networks built with *Dataset C*. The data are referred to the 7 markers group and raw data.

Dataset C	FPR	T21	OCA
Imbalanced - Normalized	4.1%	100.0%	85.2%
Imbalanced - Raw	4.3%	98.1%	84.4%
Balanced - Normalized	5.8%	100.0%	85.2%
Balanced - Raw	4.8%	100.0%	85.2%

several parameters and architectures of the ANNs, as explained in Section 3.2.7. The same structure of ANN was used in *Datasets B* and *C* and we found that the results are not consistent with *Dataset A*. Additionally, from Table 3.5 we observe that the above mentioned ANN built with *Dataset A* returns the lowest DR of the OCA.

Another objective of this study was to optimize the FPR with respect to the cost of every examination that is necessary to be done for the estimation of the risk for aneuploidy. In principle, the task is to determine the optimum number of markers required as input to the classifier. We have examined the robustness of the ANNs that are built with different combinations of input markers and we suggest two groups namely the "short" and the "long". The "short" group consisted of markers that can be extracted in one visit to the doctor. We achieve 100% detection rate of T21 at a relatively high FPR of 25%. The "long" markers group consists of three additional markers that can be extracted in another examination that measures the flow of the DV and the TF using a Doppler technique. The third marker is the NB that can be visualized during the ultra scan. The "long" markers group achieves a lower FPRs of 5%, at the same DR of 100% for T21 and >80% DR of OCA.

The population under study is not normally distributed by nature since pregnancy takes place in a certain age range which is by nature skew to the right. In addition, the large database used in this study is highly imbalanced due to the very low prevalence of aneuploidy cases in the general population, even though it contains much more trisomy cases than reported statistically. For example trisomy 21 occurs 1 in 800 pregnancies, and thus in our database we should have had less trisomy 21 cases than we actually have. Similarly trisomy 18 occurs 1 in 5000 pregnancies and thus we should again have even less cases.

One other possible reason of the non-normality distributions of the euploid may be due to the increased significance of those values that depart from normality. This is usually a case when there is a large population such as the population of the euploid cases in our database. Any method that measures the normality of a distri-

bution of large populations has high probability of rejecting the null hypothesis that the sample comes from a normal distribution.

From our experiments, see Fig. 3.5 in Section 3.3, we conclude that marker CRL does not contribute significantly to the diagnostics of the system and hence it can be ignored. The CRL is a physical distance between the crown and the rump of the fetus and the obstetrician measures it during the ultra scan. An accurate measurement of the CRL requires a specific position of the fetus and other factors that are unpredictable. Due to this fact, the distribution of the CRL values has a high standard deviation.

The balanced ANNs shown in Fig. 3.7 yield lower FPRs compared to imbalanced ANNs for the detection of the OCA while no significant difference was found for the detection of T21. Nevertheless, we suggest the use of a balanced training set for the ANNs as the MSE reduces dramatically compared to the score of the imbalanced training set. From our results, we demonstrate that the networks trained with balanced populations of euploid and aneuploid yield lower FPRs compared to the imbalanced training sets and therefore we suggest this method as a pre-processing step.

The estimation of the risk for T21 is currently done using a SMM from the combination of the markers that are explained in this chapter and there is a performance of 95% of T21 at 5% FPR. Our method outperforms the state-of-the-art method with 100% DR of the T21 at the same FPR. Additionally our method detects >80% of the OCA. From the two other non-invasive methods (SSM and cell-free fetal DNA) for the diagnosis of aneuploidy that are found in the literature, the cell-free fetal DNA test is the most promising. Several published studies report perfect separation between the euploid and aneuploid. However, there is a practical problem of this method being the cost, which is forbidding for general use. Moreover, the results are returned after some considerable time. The proposed methodology demonstrated in this chapter can be used as a first screening on the data for selecting the positively ranked cases (100% T21, 85% OCA and 5% of euploid) which will be the only ones suggested for a cell-free fetal DNA test. This will limit the overall cost for prenatal test and at the same time guarantee zero undiagnosed T21 births.

3.5 Future work

The combination of the "short" and "long" markers groups could be a two-stage procedure for a first and second screening for aneuploidy. The results of the "short" marker group are returned immediately in the personal computer of the doctor in one visit. These results assure that the negative prediction does not contain any T21. All the positive cases will be re-evaluated with the "long" group to estimate the final risk for T21 and OCA. This work will be validated and reported in a future work.

3.6 Conclusion

Diagnosis of the T21 and OCA can be effectively achieved with ANNs and a combination of biomarkers and ultrasonographic markers in the early stage of pregnancy. We have used three datasets to answer our research questions which include the detection of all the T21 cases at the lowest FPR possible, the identification of an optimal combination of input markers, the contribution of the normalized values of the data over the raw and the possible use of training sets that are consisted with balanced populations among euploid and aneuploid.

We have shown that the optimal combination of markers belongs to the "long group" which requires ultrasonographic and maternal blood examinations. Furthermore, the use of the raw data appear to be significantly more effective for the networks built with the "long group" but less effective for the "short group". Another contribution of this work is the proposed method for the data reduction of the euploid using the k-means algorithm in an attempt to create populations balanced in numbers among euploid and aneuploid. We have shown that the balanced sets appear to be more effective for training the ANNs.

In this chapter we present a system that is able to identify the entire population of T21 and the majority of the OCA such as T18, T13, Turner syndrome, and Triploidy. We have used datasets with populations that ensure statistical confidence of our results, compared to databases used in similar work of other groups. We achieved with ANNs a 100.0% detection rate of T21 and 85.2% of the OCA with FPR of 4.8%.

Submitted as:

A. Neocleous, A. Syngelaki, K. Nicolaides, and C. Schizas, "Two Stage Approach for Aneuploidy Risk Estimation Using Computational Intelligence," *Ultrasound in Obstetrics and Gynecology*.

Chapter 4

Two Stage Approach for Aneuploidy Risk Estimation

Abstract

Objective To estimate the risk for fetal chromosomal abnormalities during the first trimester of pregnancy at the lowest possible cost.

Methods As a first step, we sort the cases into two categories of "no risk" and "risk". The cases of "no risk" are no further examined, while the cases with risk are forwarded in stage 2 for further examination where we classify them in three types of risk, namely "no risk", "moderate risk" and "high risk".

Results Of a total of 37237 unknown pregnancies examined by the system, in the first stage, 13,270 euploid (33.1% FPR), 0 trisomy 21 and 47 other chromosomal abnormalities (OCA) are classified as "no risk". The remaining 6,568 cases are reassigned in stage 2 where 5,438 euploid, 0 T21 and 3 OCA are classified as "no risk", 1,085 euploid, 28 T21 and 25 OCA as "moderate risk" and 45 euploid, 61 T21 and 132 OCA as "high risk'.

Conclusion We propose a method that ensures that no T21 is classified as "no risk" in any stage. From our experiments, the best results for a TPR of 100% achieved with a FPR of 3.9%. These are euploid cases that will have to go through stages 1 and 2. Our method is non-invasive, highly effective on aneuploidy identification and it outperforms other existing statistical methods.

4.1 Introduction

Fetal chromosomal abnormalities (FCA) are diagnosed in the first trimester of pregnancy with amniocentesis (AC) or the chorionic villus sampling (CSV) tests. However, such methods are invasive and they carry risks for pregnancy complications or miscarriages (Zelig et al. 2016).

To minimize the number of invasive tests, a prenatal examination is performed to every pregnant woman. A risk for trisomy 21 (T21) is calculated with a statistical mixture model using the maternal age, the fetal nuchal translucency (NT), and the serum biochemistry (PAPP-A and β-hCG) (Nicolaides 2004, Nicolaides

et al. 1992, Spencer et al. 2000, Spencer et al. 2003). As recorded in the literature, the non-invasive prenatal test (NIPT) can identify about 90% of the T21 cases at a FPR of 3-5% (Nicolaides 2011, Maiz et al. 2009, Kagan et al. 2009). Furthermore, the diagnostic rate (DR) for T21 increases from 93 to 96% at FPR of 2.5% when the ductus venosus (DV), the tricuspid flow (TF) and the presence or absence of the fetal nasal bone (NB) is added to the system (Spencer et al. 2003). Current practice, dictates that all cases rank at high risk from the NIPT should go for invasive test in order to diagnose beyond any doubts any possible fetal chromosomal abnormalities.

Recently, another non-invasive method has shown promising results that are evaluated in big datasets (Norton et al. 2012, Norton et al. 2015, Zhang et al. 2015). The circulating free fetal DNA is isolated from the maternal blood via double centrifugation. Then, it is sequenced from the maternal plasma and distributed over the chromosomes separately for euploid, T21 and T18. Statistical techniques for classification such as Z-transform or t-test are applied to the euploidy and aneuploidy distributions of the chromosomes 21 and 18 respectively, to estimate the probability for aneuploidy of an unknown case.

The cell-free DNA test (cfDNA) has been validated for the T21, and for other chromosomal abnormalities such as trisomies 13 (T13) and 18 (T18) and some studies report perfect classification. However, some of the drawbacks are the following: First, the result of a test may take a considerable time to be available. Additionally, the maternal blood analysis requires high quality equipment and specialized doctors resulting to an increment of the general cost. From a technical point of view, there is some lack of scientific evidence on the generalization ability of this method. The majority of the papers present the results on cases that were already ranked as high-risk from the combined screening test. Therefore, 5% of the NIPT false negatives were not evaluated to test the performance of the method on statistically difficult cases. In addition to this, the study population reported in several papers is not statistically solid due to the fact that their analysis is based on small datasets (Chiu et al. 2011, Nicolaides et al. 2013, Papageorgiou et al. 2011).

In the last years, there has been an attempt to combine the two abovementioned non-invasive methods in a routine examination procedure (Nicolaides et al. 2012, Nicolaides et al. 2005). In a first stage, a risk for T21 is estimated and "high risk" cases are further examined with the cfDNA test.

The method proposed in this chapter will follow the same philosophy as above, but instead of combining statistical method and cfDNA will combine an ANN method and the cfDNA. We propose a routine examination using the markers from the prenatal examination and the use of artificial neural networks as a classifier.

The aims towards the combination of the two non-invasive methods are to optimize the number of the invasive tests at the lowest FPR possible and to minimize the cases that are suggested to perform a cfDNA test.

The proposed method is compared to the previous ones, achieves and identifies

other chromosomal abnormalities (OCA) than T21 such as T13, T18, Turner syndrome and triploidy that none of the other methods reported satisfactory results.

4.2 Methods

In this chapter we propose a two-stage approach for the estimation of the risk for aneuploidy in the first trimester of pregnancy. In Fig. 4.1, we present the overview of our proposed methodology that is based on a routine examination suggested by Nicolaides et. al (Nicolaides et al. 2012, Nicolaides et al. 2005). In the first stage, we classify an unknown case into "no risk" or "high risk" using a cut-off value (threshold) that is applied to the output of our classifier. We use four markers from a standard pre-natal examination. The maternal age contributes significantly to the discrimination of the euploid and aneuploid classes and it is used as a marker for both stages. The serum biochemistry markers, the PAPP-A and the β-hCG are derived from the maternal blood and the results are observed within few minutes using automated machines (DELFIA Xpress system, PerkinElmer Life and Analytical Sciences, Waltham, MA). The nuchal translucency (NT) is a subcutaneous collection of fluid behind the fetal neck and it is measured by transabdominal ultrasonography.

All the "high-risk" cases from the first stage are reassigned in the second stage. We use two cut-off values that we apply to the output of the system in stage 2, to classify the unknown case as "no risk", "moderate risk" or "high risk". In this stage we include three additional markers, the DV, TF and NB. The DV and the TF are measured from a special ultrasonographic scan, using a pulsed wave doppler technique. The NB is observed during the first scan. We propose that the "moderate risk" cases are suggested for cfDNA test while the "high risk" to perform an invasive test.

In the field of biological informatics, the use of computational intelligence techniques such as artificial neural networks (ANNs), support vector machines (SVM) and others have brought interesting results in the prediction of serious diseases such as cancer, parkinson and others (Schnorrenberg et al. 1997, Schizas and Pattichis 1997, Neves et al. 2015, Er et al. 2016, Yoo et al. 2016). The ANNs have the advantage of "learning" by examples that are presented one by one in a similar way to the brain process of learning. This ability of learning, differentiates them from the standard statistical methods where the distributions, the mean and standard deviations are significant.

The feed forward (FF) ANNs are consisted of a number of layers that are connected to each other. The first layer is called input layer and it contains as many nodes (neurons) as the input parameters. The last layer is called output layer and it

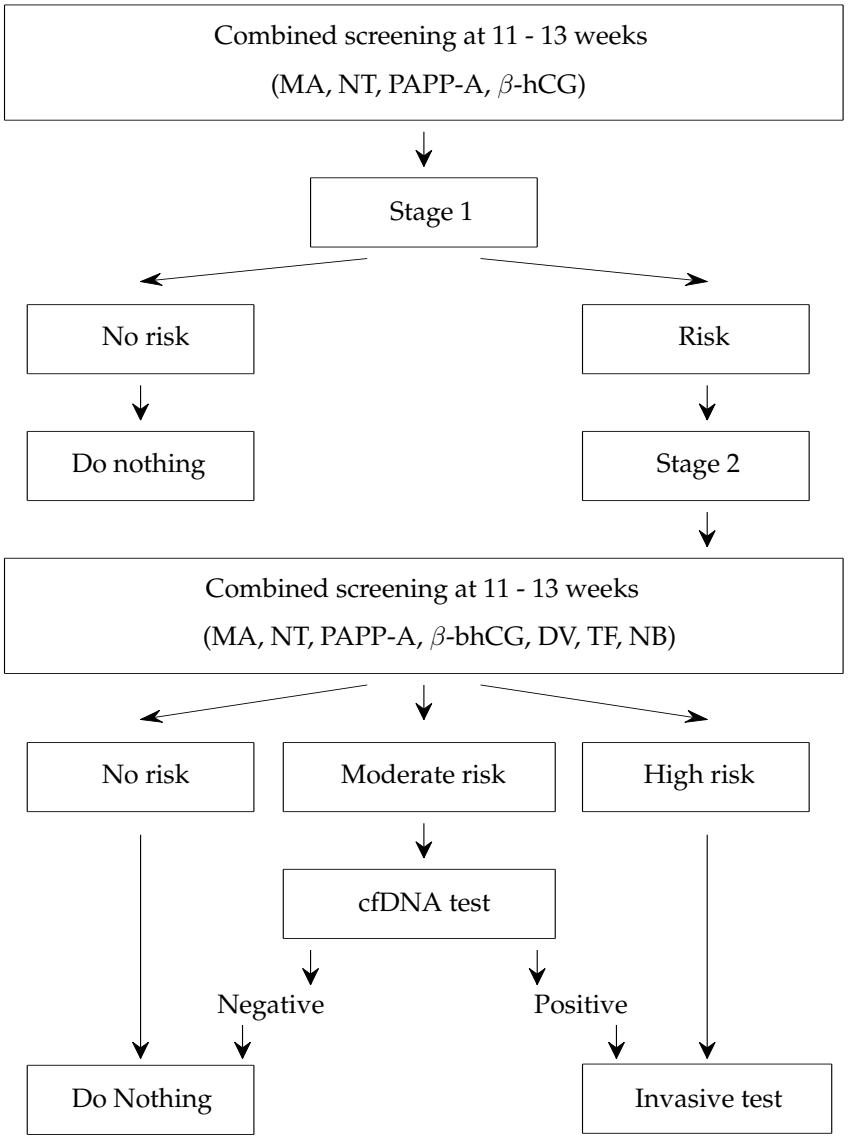

Figure 4.1: Overview of the proposed methodology. Every case is suggested to perform the first stage of the prenatal examination for estimating the risk for fetal aneuploidy. The "high risk" cases are reassigned in stage 2 that are finally classified in "no risk" and continue the pregnancy, "moderate risk" and suggested to perform the cfDNA test or "high risk" and suggested for an invasive test for reaching diagnosis.

contains one node. The rest of the layers that are between the first and the last, are called hidden layers. A typical ANN architecture is consisted of one or two hidden layers. In every connection, there is a weight and an activation function that represent the process in the synapses of the biological brain cells. The weights are optimized during the training procedure by presenting all the examples several times and calculating the error, which consequently is used for adjusting the weights through a learning algorithm. The number of repetitions is a parameter called "epochs". The number of layers and nodes, the activation functions and the epochs are parameters that are pre-defined.

An unknown case is evaluated by an ANN by presenting the parameters to the input layer. This information is passed through the layers by applying the appropriate weights and the transfer functions. The output value of the ANN in the last layer takes values in the range between 0 and 1 due to the sigmoid transfer function. To classify a case into negative or positive, we use a cut-off point. The output values that exceed this point are considered positive.

4.3 Results

For our experiments, we used three different datasets that were provided by the Fetal Medicine Foundation at different periods of time. The total number of euploids is 122362 while the aneuploids are 967, including 684 T21, 141 T18, 46 T13, 38 triploidy and 58 Turner cases (table 4.1).

To ensure the robustness of the ANN models, we applied a three-fold cross validation procedure. We first split our datasets into a training set of 70% of the total populations and a test set (30%) that is kept unknown to the training phase. The populations of the cases that consist the cross validation sets (training and tests) for the three datasets A, B and C are shown in table 4.2.

To identify the optimal parameters of a neural network is generally done empirically (chapter 2). The identification of the ANN that yield the best results, we used the training set of the dataset A that we call the "training dataset" to perform a grid search of 24 ANNs, by systematically changing different parameters. Then, we performed our experiments in datasets B and C that we call the "testing datasets" with the same parameters of the ANN from dataset A. In table 4.3, we present the results of the grid search. We choose to use the parameters of the FF #18 (table 4.3), to build the neural networks for our experiments using the datasets B and C .

In Tables 4.4, 4.5, 4.6 and 4.7 we present the results of the "testing datasets" separately. The results of the cases in the test sets that were ranked with risk in stage 1 are shown in Table 4.4. These are the cases that are suggested a further examination and send for testing in stage 2. In tables 4.5, 4.6 and 4.7 we present the results of stage 2 for the three categories "No risk", "Moderate risk" and "High

Table 4.1: Euploid and Aneuploid Populations of the Three Datasets Used.

Dataset	Euploid	T21	T18	T13	Triploidy	Turner
A	50517	408	39	14	10	13
B	29790	124	42	10	14	19
C	42055	152	60	22	14	26

Table 4.2: Cross Validation. Training and Validation Sets for the Three Datasets Used.

	Training		Validation		
Dataset	Euploid	T21	Euploid	T21	OCA
A	33619	279	16898	129	76
B	20782	100	9008	37	85
C	31225	100	10830	52	122

risk".

In order to evaluate both "testing datasets" using a unique cut-off point, we concatenate the outputs of the test sets of datasets B and C and we present our results in Figs. 4.2, 4.3 and 4.4. In Fig. 4.2, we present the ROC curves of the ANNs built in stage 1 with dashed and stage 2 in solid lines. These curves are calculated using different values of cut-off points that is applied to classify an unknown case as euploid or aneuploid. For every case in the test sets, we compare the result of the classification with the real karyotype to determine the result as true or false poitive/negative. We calculate our results in terms of sensitivity and specificity:

$$sensitivity = \frac{TP}{TP + FN} \tag{4.1}$$

$$specificity = \frac{TN}{TN + FP} \tag{4.2}$$

The true positive (TP) is the total number of the aneuploid cases that are classified positive and false negative (FN) the ones that are classified as negative. Respectively, the euploid cases that are classified as negative are called true negative (TN) and as positive, false positive (FP).

In Fig. 4.3, we present analytically the classification results of the testing datasets. From the total number of 20134 cases that were validated in stage 1, 33.1% of the

Table 4.3: Architecture Structures of the Neural Networks built and the results of the grid search.

Network	Neurons #	Activation	FPR	T21 #	OCA
FF 01	5	Logistic	96.0%	100.0%	36.8%
FF 02	10	Logistic	93.9%	100.0%	47.4%
FF 03	15	Logistic	93.5%	100.0%	51.3%
FF 04	20	Logistic	94.4%	100.0%	43.4%
FF 05	25	Logistic	95.8%	100.0%	36.8%
FF 06	30	Logistic	92.0%	100.0%	52.6%
FF 07	35	Logistic	95.3%	100.0%	23.7%
FF 08	40	Logistic	93.1%	100.0%	43.4%
FF 09	45	Logistic	95.0%	100.0%	19.7%
FF 10	50	Logistic	93.3%	100.0%	43.4%
FF 11	55	Logistic	95.9%	100.0%	19.7%
FF 12	60	Logistic	96.3%	100.0%	30.3%
FF 13	5	Tanh	96.8%	100.0%	25.0%
FF 14	10	Tanh	94.0%	100.0%	35.5%
FF 15	15	Tanh	95.9%	100.0%	39.5%
FF 16	20	Tanh	95.2%	100.0%	40.8%
FF 17	25	Tanh	95.6%	100.0%	22.4%
FF 18	30	Tanh	96.3%	100.0%	32.9%
FF 19	35	Tanh	94.7%	100.0%	28.9%
FF 20	40	Tanh	93.0%	100.0%	39.5%
FF 21	45	Tanh	93.7%	100.0%	32.9%
FF 22	50	Tanh	96.6%	100.0%	32.9%
FF 23	55	Tanh	95.8%	100.0%	32.9%
FF 24	60	Tanh	94.9%	100.0%	30.3%

euploid were classified as positive together with the 100% of T21 and 77.3% of the OCA that are reassigned in stage 2.

Table 4.4: Results of stage 1 for the "Risk" category.

Dataset	Euploid	T21	OCA
B	26.9%	100%	82.4%
C	31.7%	100%	72.1%

Table 4.5: Results of stage 2: "No risk"

Dataset	Euploid	T21	OCA
B	93.7%	0.0%	2.4%
C	96.1%	0.0%	0.0%

Table 4.6: Results of stage 2: "Moderate risk"

Dataset	Euploid	T21	OCA
B	6.2%	35.1%	16.5%
C	3.7%	25.0%	13.9%

Table 4.7: Results of stage 2: "High risk"

Dataset	Euploid	T21	OCA
B	0.2%	64.9%	77.1%
C	0.2%	75.0%	58.2%

Each of the positive cases from stage 1 is classified from stage 2 into one of the three categories of "no risk", "moderate risk" or "high risk" based on the ANN output value and two cut-off points. We set the lower cut-off point to the 11% of risk to separate the "no risk" with the "moderate risk". The cut-off point for the "high risk" is set to 95% of risk.

From the total number of 6568 cases that were ranked positive in stage 1, the 82.8% of the euploid, 0% of the T21 and 1.8% of the OCA (1.4% of the total number of OCA) is classified as "no risk" crating a total FPR of 5.7%. In the "moderate risk" category, 16.5% of the euploid (5.5% of the total), 31.5% of the T21 and 15.6% of the OCA (12.1% of the total) are suggested to perform a cfDNA test. In the "high risk" category, 0.7% of the euploid (0.2% of the total), 68.5% of the T21 and 82.5% of the OCA (63.8% of the total) are suggested to perform an invasive test.

In Fig. 4.4, we present the outputs of the 13270 euploid cases (upper plot) and the

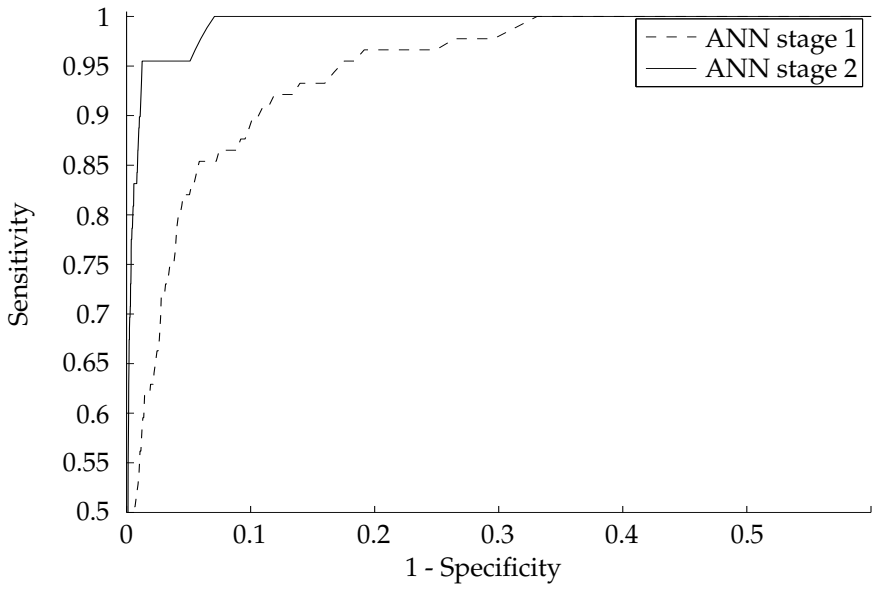

Figure 4.2: ROC curves of the neural network outputs of stage 1 (dashed line) and stage 2 (solid line). The sensitivity and the specificity were calculate for different values of cut-off points and are placed in a two-dimensional plot.

outputs of the 249 chromosomal abnormalities in the lower figure that are ranked positive in stage 1. In the y-axis of each plot, we show the percentage of the risk for every case that is based on the output of the ANN in stage 2. The continuous line represents the cut-of point that separates the categories "no risk" and "moderate risk". The dashed line represents the cut-of point that classifies the cases as "high risk".

4.4 Discussion

In this work we present a two-stage approach for the estimation of the risk for aneuploidy. In both stages, we make sure by adjusting the threshold accordingly that no trisomy 21 cases are classified as euploid (i.e. 100% TPR). At the same time, we focus our studies to minimize the FPR to the lowest possible. We have validated our results using two test sets from datasets B and C of a total number of 19838 euploid, 89 T21 and 207 OCA cases.

In stage 2, we use seven markers (MA, PAPP-A, β-hCG, NT, DV, TF and NB) as inputs to the ANN. The important contribution of the last three parameters is indicated by the steeper shape of the associate curve in Fig. 4.2. From the ROC

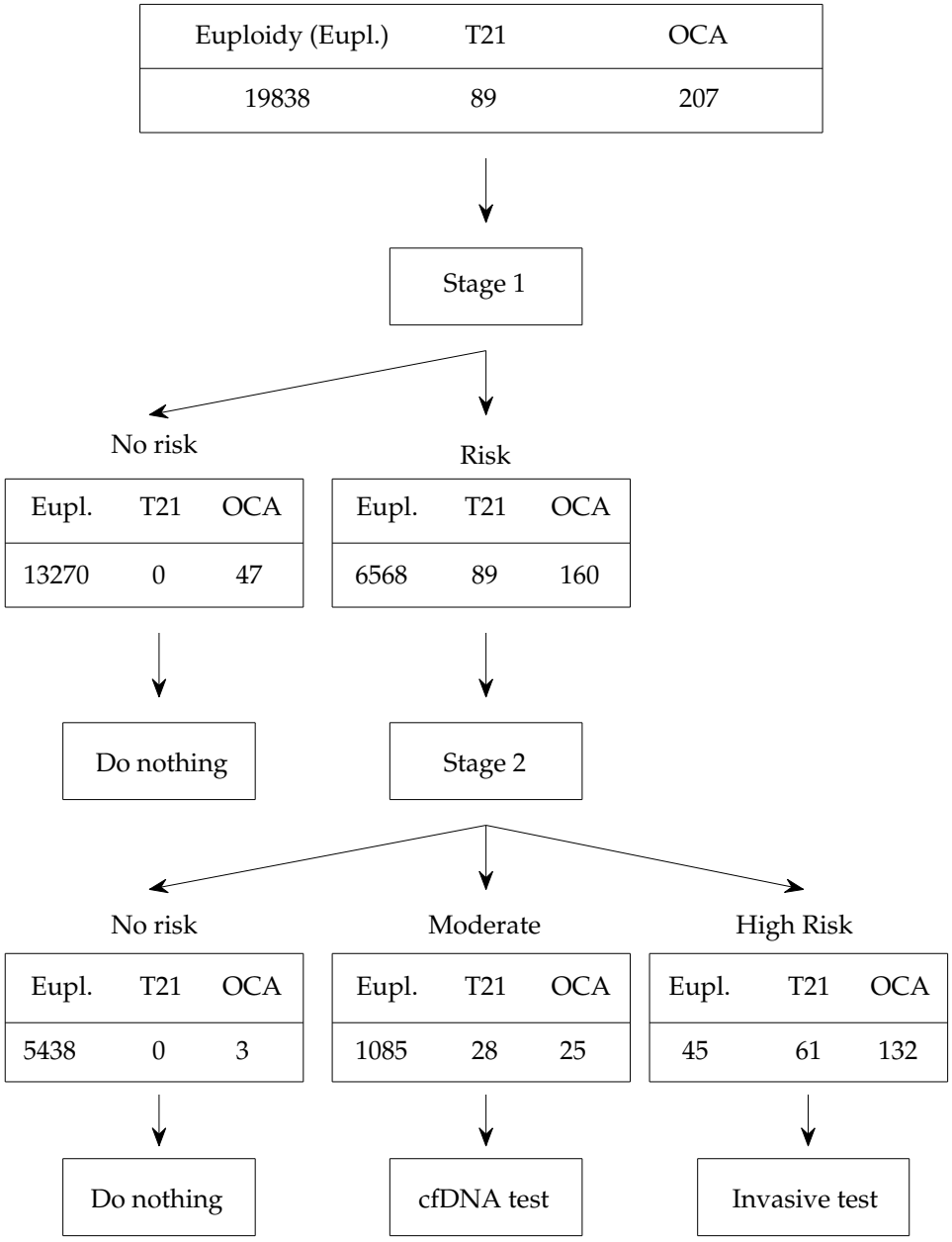

Figure 4.3: The results of the testing datasets A and B for stages 1 and 2.

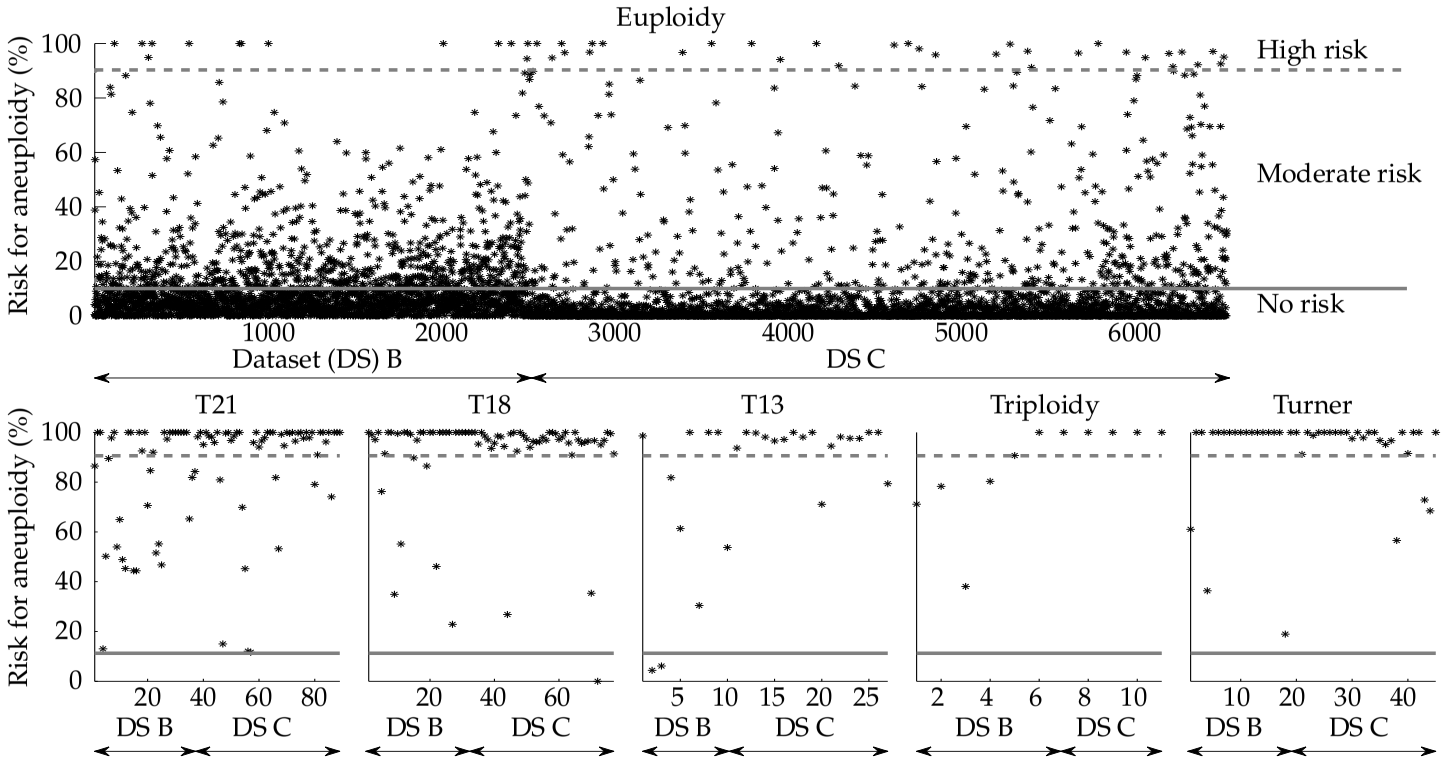

Figure 4.4: Distribution of the risk for aneuploidy for the cases that were ranked positive in stage 1, including the 33.1% of the euploid (upper plot) and the five chromosomal abnormalities in the lower plots. Below every plot we annotate with double arrow the cases that belong to datasets B and C. The lower constant lines indicate the cut-off point that separates the "low risk" with the "moderate risk" categories. The up4per dashed lines represent the cut-off point for the "high risk" category.

curves, it is shown that the ANNs built with seven markers perform significantly better than the ANNs built with four markers.

We report higher classification results than the state-of-the-art statistical mixture model that is currently used as a classifier. In order to make an accurate comparison between our method and the NIPT test, we compare the 95% DR for T21 that is reported in the literature at the 5% FPR. We use a cut-off point at the 45% of the risk, as explained in the section Methods, and achieved 94.2% and 79.5% DRs for T21 and the OCA respectively at FPR of 1.2%.

In addition to the diagnosis of T21, our method achieves high accuracy on the detection of the OCA, as shown in Fig. 4.4. Below every plot, we annotate with double arrows the cases that belong to datasets A and B. From the upper plot, we observe that dataset C has better performance on the euploid cases. The same test dataset identified 100% of the triploidy cases. From Fig. 4.4, we conclude that all the T21, the triploidy and the Turner cases are ranked as "moderate risk" or "high risk" with a cut-off point of 11%.

The FPR in the first stage kept relatively high for accommodating all the T21 and 77.3% of the OCA. In the second stage, we estimate another risk based on the results of the ductus venosus, the tricuspid flow and the nasal bone. By including the ductus venosus, the tricuspid flow and the nasal bone, the FPR is reduced from 33.1% to 5.7%.

In the literature, the cfDNA test has shown in many studies to be appropriate for the estimation of the risk for T21 and T18 reporting almost perfect classification. However, in the majority of these studies, the population that is used for testing has already ranked as "high risk" from a previous non-invasive method. Therefore, there is no statistical evidence about the 5% of the false negative T21 and the 5% FPR that is left outside the test set. In addition to this, since the classification is done using standard statistical classifiers such as t-test and z-scores, the results should be cross-validated to ensure that the models are robust to pattern changes. Two other drawbacks of the cfDNA test are a) the complexity of the method and the fact that in about 1.0% of the population the test returns no results.

Our proposed methodology has the potential to be used in a real time application in medical centers, since it returns immediate results during a regular visit of the pregnant woman. Some of the drawbacks of our method are that it does not classify perfectly the euploid from the aneuploid since we still report a cumulative FPR of 5.7%. Additionally, more work needs to be done for improving the DR of the OCA.

In this work, we make sure that no T21 will be born unexpectedly, while we reduce the cases that nowadays unnecessarily undergo for the cfDNA test.

For future work, we will focus our research to build models that will associate the risk for aneuploidy with pre-eclampsia and other pregnancy complications. We currently have preliminary results that we aim to publish in high impact academic journal.

Part II

Signal Processing and Data Mining in Computational Ethnomusicology

Prepared for submission:

A. Neocleous, G. Azzopardi, C. N. Schizas and Nicolai Petkov, "COSFIRE flters for 1-D pattern recognition with application to musical signals" *EURASIP journal of signal processing and music*.

Chapter 5

Melodic similarity using COSFIRE filters

Abstract

The identification of repeating patterns in time signals - also known as motifs - has been a challenge for many years in the area of digital signal processing. For instance, in music, existing algorithms that detect similar motifs typically require multi-dimensional features such as chromagrams, similarity matrices and the use of the computationally expensive dynamic time warping (DTW) algorithm. We propose a time-scale and transposition invariant method for identifying repeating motifs and demonstrate its effectiveness in musical signals. We adapt the COSFIRE approach, which has been found effective in 2D signals, to 1D signals. Our method has a computational time complexity substantially lower than DTW and achieves better effectiveness compared to DTW, Symbolic Aggregate Approximation and cross correlation results. The proposed 1D COSFIRE approach is highly effective and efficient for extracting a symbolic representation of the melodies in a given song. Additionally, it is conceptually simple and versatile, in that it can be applied to any 1D signal.

5.1 Introduction

The identification of similar or identical sub-sequences that appear in different positions in 1D time series, has been a task for research in the fields of digital signal processing and pattern recognition for the last two decades. Such subsequences are typically called motifs. Computational methods have been applied in medical and biological signals (Abe and Yamaguchi 2005) and for tasks such as weather prediction (McGovern et al. 2007), speech recognition (Payne 2006) and music (Müller et al. 2011, Eronen 2007). In this work we propose a novel computational method to identify similar 1D patterns and demonstrate its effectiveness in musical signals. In the following, we therefore focus on the problems and literature relating to music pattern recognition.

In music theory, a motif can be defined as the smallest melody with an important thematic identity. Such a pattern can be a defining feature of a musical piece − think for instance of the sequence A-G-F-E-D-C#-D in Toccata and Fugue in D minor by J.

S. Bach − and thus be of central importance in the context of intellectual property. The verse of a song is an example of such a repeating pattern in popular music.

The term *melody* can be primarily defined as a sequence of musical tones and silences of specific duration over time. In a first approximation, a melody can be described mathematically by a sequence of fundamental frequencies over time called pitch track, derived from the audio signal. The term *pitch* refers to a perceptual property of a musical tone and is closely related to the fundamental frequency of that tone. The complexity of the relation between the pitch and the fundamental frequency has been described in (Butler 1989, Howell et al. 1991).

The identification of motifs leads to an explicit understanding of the musical structure of a piece. The detection of such patterns is an essential step in musicological analysis. In Fig. 5.1 we illustrate an example from a data set that we compiled for this work[1]. Fig. 5.1a shows the values of the microphone signal of a song. In Fig. 5.1b, we show a representation of the melody short-term fundamental frequency over time. In this case four motifs are present. The boundaries of the motifs are indicated with vertical lines. Motif 3 is a repetition of motif 1 and they are both indicated with the letter "A". Similarly, the letter "B" is given to the motif 2 and motif 4. The musical structure of the song in Fig. 5.1 has the form "A" - "B".

In this work we focus on monophonic songs of Eastern Mediterranean folk music. There are various differences between popular and folk music which have to be taken into account in computational analysis. Popular music is typically recorded in professional studios with high quality equipment and professional musicians. In folk music, non-professional performers may perform singing with rhythmic variations and pitch drifts. However, generally, it is difficult to determine whether a performer (professional or non-professional) is being expressive or hiding a mistake. In addition, recordings of folk music are often made in noisy environments. From a technical point of view, the majority of the methods for similar problems are using Musical Instrument Digital Interface (MIDI) information in their methodology. Folk music lacks from transcriptions since many times the composer is not known and therefore similar melodies can be found in different songs. Therefore, in the computational analysis of folk music it is preferable that the MIDI information is avoided.

In this chapter we propose a method to identify repeating patterns using a single feature, the short-time fundamental frequency, as a function of time. This method can be applied to monophonic (i.e. with a single melody) vocal folk tunes. It is based on a modification of the COSFIRE (Combination Of Shifted Filter Responses) approach (Azzopardi and Petkov 2013), a technique that has been introduced for trainable visual pattern recognition. So far, the COSFIRE method has been applied in two-dimensional signals. In this work, we adapt the COSFIRE approach to 1D

[1]Data set can be downloaded from https://www.cs.ucy.ac.cy/projects/folk/

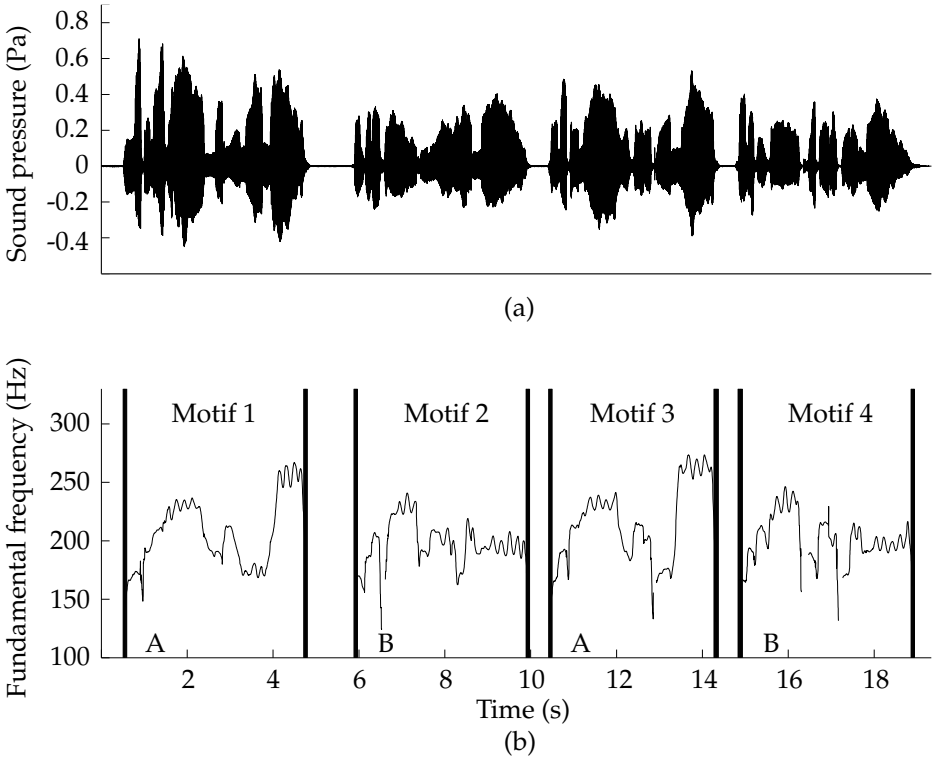

Figure 5.1: Audio signal represented as (a) obtained by a microphone and as (b) a sequence of fundamental frequencies coming from a short-term frequency analysis. In this example, four motifs are present, whose starting and ending positions are indicated by the vertical lines.

signals. We compare our results with other distance metrics that are widely used in pattern recognition of 1D problems.

5.1.1 Related work

The majority of the applications in Music Information Retrieval (MIR) require the extraction of low-level features derived from the raw audio signal. There has been a significant amount of work in the research of audio feature representation (Peeters and Rodet 2004, Mckinney and Breebaart 2003).

Depending on the application, different groups of features may be used. Temporal features include the autocorrelation coefficients or the note duration. Spectral information can be described by Fourier Transform (FT), spectral spread, spectral centroid, Mel frequency cepstral coefficients (Zheng et al. 2001). These features are usually used for other tasks such as instrument identification (Benetos

et al. 2006, Giannoulis and Klapuri 2013), musical genre classification (Tzanetakis and Essl 2001, Tzanetakis and Cook 2002) speech/music identification (Ramalingam and Dhanalakshmi 2014, C and Chang 2012, Neocleous et al. 2014b), speech processing and recognition (Wu and Li 2013, Saeidi et al. 2012, Lee et al. 2014, Ishizuka and Nakatani 2006, Zade et al. 2006) and others.

In computational analysis of music structure, several methods have been proposed for the automatic identification and segmentation of important musical parts. These are often called audio thumbnailing (Bartsch and Wakefield 2005). Such parts include the introduction, the verse and the chorus. Other methods require note segmentation and the identification of repeating patterns is eventually achieved by creating a symbolic representation of the melody and the use of chromagrams (Bartsch and Wakefield 2005, Gómez 2006, Goto 2006, Müller et al. 2011). The studies in (Qi et al. 2007, Sandler and Aucouturier 2001) include audio segmentation using self-similarity measures and hidden Markov models.

In (Bartsch and Wakefield 2005) the authors propose a method for audio thumbnailing of popular music. Their method includes a frame-based audio segmentation where each audio frame is described with a set of features. Then, similarity matrices of the extracted features are computed and used to identify motif repetitions in the entire signal. The most common feature used for audio thumbnailing is the 12-dimensional chromagram proposed in (Gómez 2006).

Similar approaches to audio thumbnailing and automated identification of repeating musical aspects are described in (Goto 2006, Müller et al. 2011, Eronen 2007). The majority of those papers report the use of Dynamic Time Warping (DTW) in their methodologies. The extraction of chromagrams, the computation of similarity matrices and the DTW result in a complex system with a high computational cost. The use of low-level features for audio thumbnailing is explained in (Peiszer 2007, Aucouturier and Sandler 2002).

Other studies extract information from MIDI scores and correspond the musical sequence with letters of the alphabet (Crochemore et al. 2001, Hillewaere et al. 2009). Examples of such information include the duration and the fundamental frequency of the notes, the frequency difference between the previous, current and subsequent notes, temporal onsets, and others. The musical patterns and repetitions are identified with the use of string methods such as n-gram models.

In the last two decades, the query-by-humming (QBH) method has been widely used for melodic similarity (Song et al. 2002, Hu et al. 2003, Zhu and Shasha 2003, Dannenberg et al. 2004, Dannenberg and Hu 2004, Ryynänen and Klapuri 2008, Huq et al. 2010, Kotsifakos et al. 2012). Commercial systems such as Shazam, MiDomi, musipedia and SoundHound as well as academic systems such as Tunebot are examples of applications that use the QBH method. The procedure is as follow: the audio signal is first converted into a MIDI information. This step requires pitch track extraction and note segmentation that are still a challenging problem in the

MIR. Then, the symbolic representation of the input signal is compared with a large set of songs in a database. In each comparison, a similarity value is returned. The best match between the input signal and a song in the database is chosen as the pair with the lowest similarity value.

The constraint of the methodologies discussed above, is that the MIDI scores have to be known. In the majority of the folk songs a MIDI score is not available, and therefore these methods cannot be applied. Another limitation of this methodology is the inability of analysing non-Western music. This arises from the fact that many cultures do not have a written language for describing their music, therefore a protocol such as MIDI is not available.

Another approach that is reported in the literature for motif recognition and classification is the use of wavelets (Velarde et al. 2013, Jeon et al. 2009). The melody is first represented with a single-scale signal that is derived from the continuous Haar wavelet transform. Then, wavelet coefficients are used to form feature vectors and processed for classification purposes with standard machine learning tools.

The Symbolic Aggregate Approximation (SAX) is a method that converts a time series into a symbolic representation (Keogh et al. 2006). Its major novelty that differentiates it with other established methods is that it allows for dimensionality reduction and the ability of applying distance measures on the symbolic representation. In (Lin et al. 2007) it is shown that this method had been applied for several types of signals such as the analysis of protein unfolding data (Ferreira et al. 2006), meteorological data (McGovern et al. 2006), telemedicine time series (Duchene et al. 2005) and others.

While various methods for automatic segmentation into meaningful musical sections have been proposed for Western and popular music, little work has been done on folk music (Dutta and Murthy 2014, Ishwar et al. 2013, Ross et al. 2012, Volk and Van Kranenburg 2012, Rao et al. 2014). However, the data sets used in the aforementioned studies are not available, or they are encrypted with the MIDI protocol that is not appropriate for the application at hand.

There are essential differences between these two categories. The structure of the songs includes well defined repetitions with constant tempo and rhythm. In contrast, folk music is produced by non-professional performers and most of the times these artists do not have formal musical education. This results in non-professional performances, such as singing out of tune or rhythmical variances throughout the song. Furthermore, the environmental conditions include low quality recordings in noisy public places. Existing methods for the automatic identification of repeating patterns are not equipped to deal with the mentioned aspects of folk music.

This chapter is structured as follows. In Section 5.2 we present our proposed method, while in Section 5.3 we demonstrate its effectiveness and compare it with existing methods on a new data set of monophonic folk songs. In Section 5.4 we discuss certain aspects of the proposed approach and then we draw conclusions in

Section 5.5.

5.2 Methods

5.2.1 Overview

In Fig. 5.2 we illustrate the whole pipeline of our method. It consists of a configuration and an application stage.

In the configuration, we first divide the audio signal into overlapping frames of 30ms in length with 3ms overlap. For each such a frame, we extract the fundamental frequency (Fig. 5.1b) using the YIN algorithm (de Cheveigné and Kawahara 2002) as explained further in Section 5.2.2. Then, we manually segment the audio signal into motifs as shown in Fig. 5.1b. After the segmentation is done, we take the first motif which we denote by the letter "A" and use it to configure a set of COSFIRE filters that describes its properties.

In the application stage we apply the configured set of COSFIRE filters to the remaining motifs of the audio signal. In order to come to a single value for every motif, we combine the responses of the set of COSFIRE filters by geometric mean. We label with the letter "A" every motif that evokes a combined response that is higher than a threshold, which is set experimentally. In the next iteration, we denote by "B" the next motif that has not been labelled and use it to configure a new set of COSFIRE filters. We then give the label "B" to motifs which exhibit responses higher than some threshold. We repeat this procedure until all motifs in a given audio signal are labelled.

These steps are explained in more detail in the sections below.

5.2.2 Fundamental Frequency Extraction

The fundamental frequency is closely related to the pitch of a musical note. The melody can be represented and modelled as the pitch in a function of time. The fundamental frequency estimation for musical signals has been an important and challenging task in the research of music information retrieval. Several algorithms and techniques have been proposed, some of them based on the time domain (Medan et al. 1991, Talkin 1995, de Cheveigné and Kawahara 2002) and others based on the frequency domain (Klapuri 2004, Dorken and Nawab 1994). We employed the widely used YIN algorithm proposed by Cheveign and Kawahara (de Cheveigné and Kawahara 2002) for the extraction of the fundamental frequency vector. The YIN algorithm takes as input the raw observations of the sound pressure and outputs a vector with fundamental frequency candidates for each audio frame. The algorithm is based on the autocorrelation function (ACF) of the audio signal where a number of modifications and corrections are applied after the autocorrelation, as

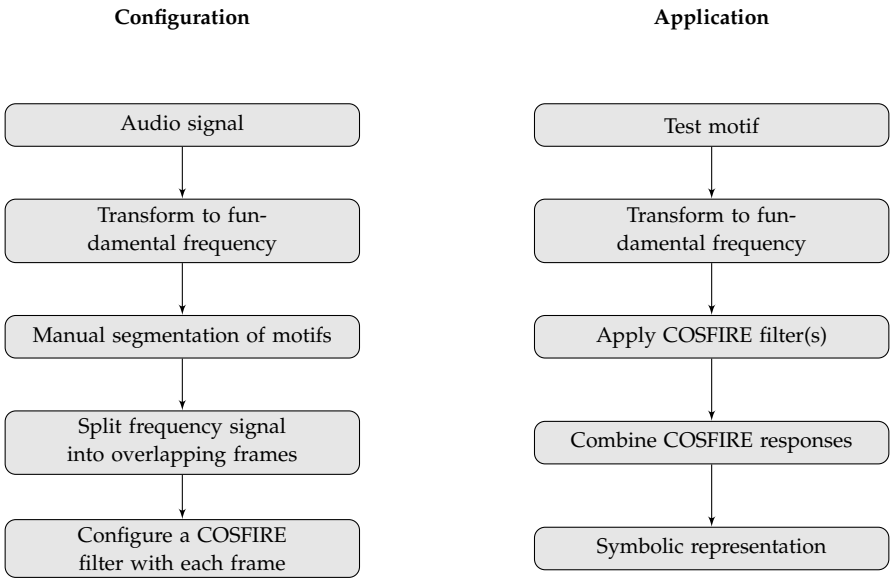

Figure 5.2: The main steps of the proposed methodology.

reported in (de Cheveigné and Kawahara 2002). The algorithm uses a frame-based approach and it gives the fundamental frequency candidates for a sequence of overlapped audio frames. For our analysis we used 1470 bins with a window length of 30ms and 128 bins for hop size (3ms).

5.2.3 Configuration and response of a COSFIRE filter

A COSFIRE filter is configured by the values at certain positions in a given training motif (also referred to as a prototype signal). We consider a regularly spaced set of n points along the given signal where the midpoint of this set lies on the center of the prototype. In Fig. 5.3a, we present a synthetic signal consisting of five different patterns. The first pattern shows a sinusoidal signal that we use as a prototype along which we consider seven points centered around the zero crossing. We describe each point i by a pair (f_i, t_i), where f_i is the value of the signal at time point t_i with respect to the center of the filter support (labeled by the * marker). We denote by P_c a COSFIRE filter that is defined as a set of such pairs:

$$P_c = \{(f_i, \rho_i) \mid i = 1 \dots n\} \tag{5.1}$$

where $\rho_i = \delta(i - (n+1)/2)$, n is the total number of considered time points and δ is the length of the interval between the time points.

For the given prototype shown in Fig. 5.3a we use (n =) 7 time points uniformly distributed in intervals of (δ =) 10 time points to configure a COSFIRE filter with parameter values specified by the pairs in the following set:

$$P_c = \left\{ \begin{array}{l} (f_1 = 0.87,\ \rho_1 = \text{-}30), \\ (f_2 = 0.95,\ \rho_2 = \text{-}20), \\ (f_3 = 0.58,\ \rho_3 = \text{-}10), \\ (f_4 = 0,\ \ \rho_4 = 0), \\ (f_5 = \text{-}0.58,\ \rho_5 = 10), \\ (f_6 = \text{-}0.95,\ \rho_6 = 20), \\ (f_7 = \text{-}0.87,\ \rho_7 = 30) \end{array} \right\}$$

The response of a COSFIRE filter is computed by first comparing the values in a given test signal to the preferred ones defined by P_c. Thus, we obtain a similarity value for each pair in P_c. Then we use a smoothing function to allow for some temporal tolerance and finally we combine all the smoothed similarity values with geometric mean.

Similarity measure

For a given point in time t of a test signal T we use a Gaussian kernel function to compute a similarity value for each pair in set P_c that defines a COSFIRE filter:

$$D_i(t) = \exp^{-\frac{(f_i - T_{t+\rho_i})^2}{2\sigma^2}}, \quad \sigma = \sigma_0 + \alpha(|\rho_i|) \tag{5.2}$$

where f_i is the preferred value of the i-th pair in set P_c, and $T_{t+\rho_i}$ is the corresponding value in the concerned neighbourhood of a signal T at time t. The standard deviation σ of the Gaussian kernel function increases linearly with increasing distance from the center of the filter. In this way we allow more tolerance to the values of time points that are on the periphery of the support of the filter than those that are closer to the support center.

In Fig. 5.4 we present an example of the prototype explained in Section 5.2.3 that comprises seven points. We use vertical lines to indicate the seven positions of the time points that are considered in this COSFIRE filter. Their heights indicate the standard deviation σ that we use in Eq. 5.2 for σ_0 = 0.1 and α = 0.9.

Temporal tolerance

In order to allow for temporal tolerance we smooth the similarity signals D_i of each pair of a COSFIRE filter. We denote by $S'_i(t)$ a smoothing function that we com-

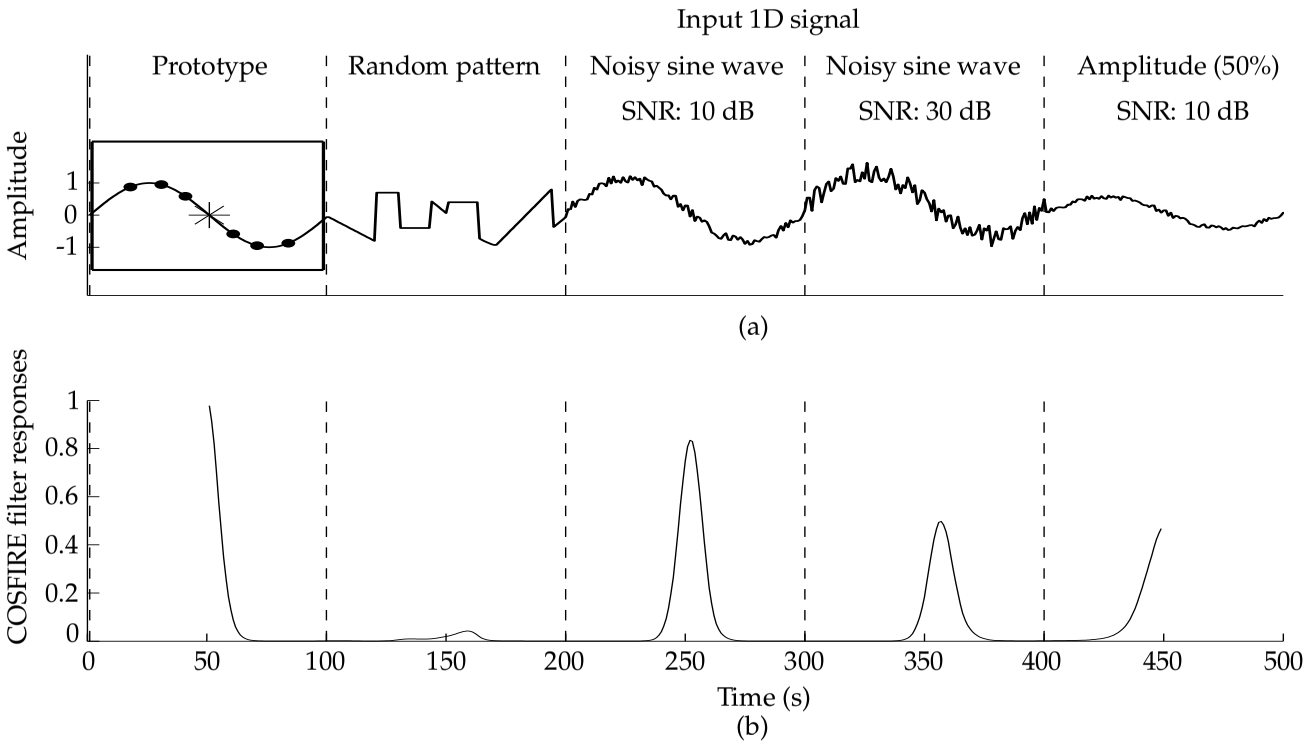

Figure 5.3: (a) Input 1D signal. The enframed sinusoidal signal is used as a prototype to configure a COSFIRE filter. (b) The responses of the filter to the input signal. The highest response is achieved for the prototype signal, but strong responses are also recorded at the center of similar but noisy patterns. Negligible response is achieved for the random pattern.

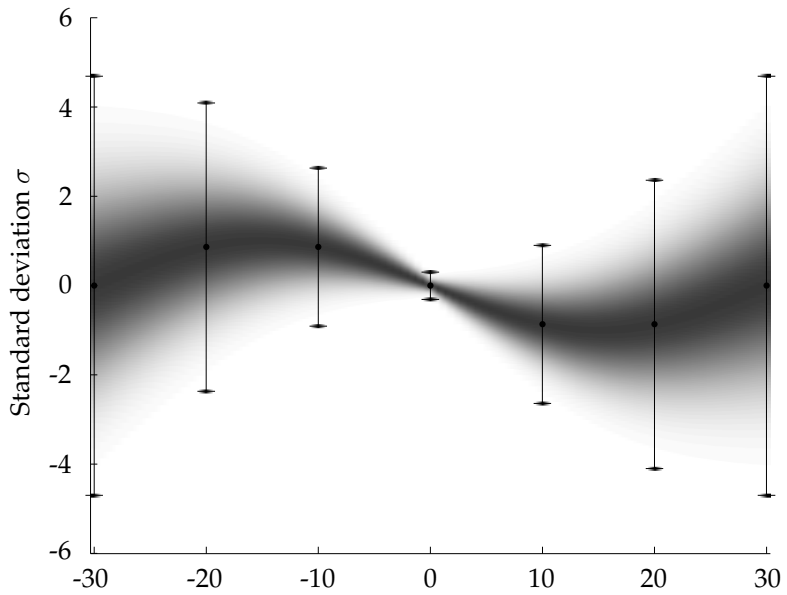

Figure 5.4: The value of the standard deviation increases for each point i away from the support center of the filter.

pute by taking the maximum value of the neighbourhood of the resulting similarity signals weighted by a Gaussian function with a standard deviation σ':

$$S'_i(t) = \max_{-3\sigma' \leqslant j \leqslant 3\sigma'} \{D_i(t+j) \exp^{-\frac{j^2}{2\sigma'^2}}\} \tag{5.3}$$

In Fig. 5.5a we use the same prototype that we used in the example in Fig. 5.3a to configure a COSFIRE filter with $n = 100$, $\delta = 1$, $\sigma_0 = 0.01$ and $\alpha = 0.001$. The test signal is a noisy sinusoidal signal similar to the prototype but with a lower frequency. We use this example to illustrate the effectiveness of the smoothing function to achieve temporal tolerance. The image in Fig. 5.5b is a matrix of (n =) 100 similarity signals, one for each pair in the set P_c of the concerned COSFIRE filter. Due to the boundary effect, we do not compute similarity values in the beginning and end parts of the input signal. In Fig. 5.5c we show the smoothed similarity signals.

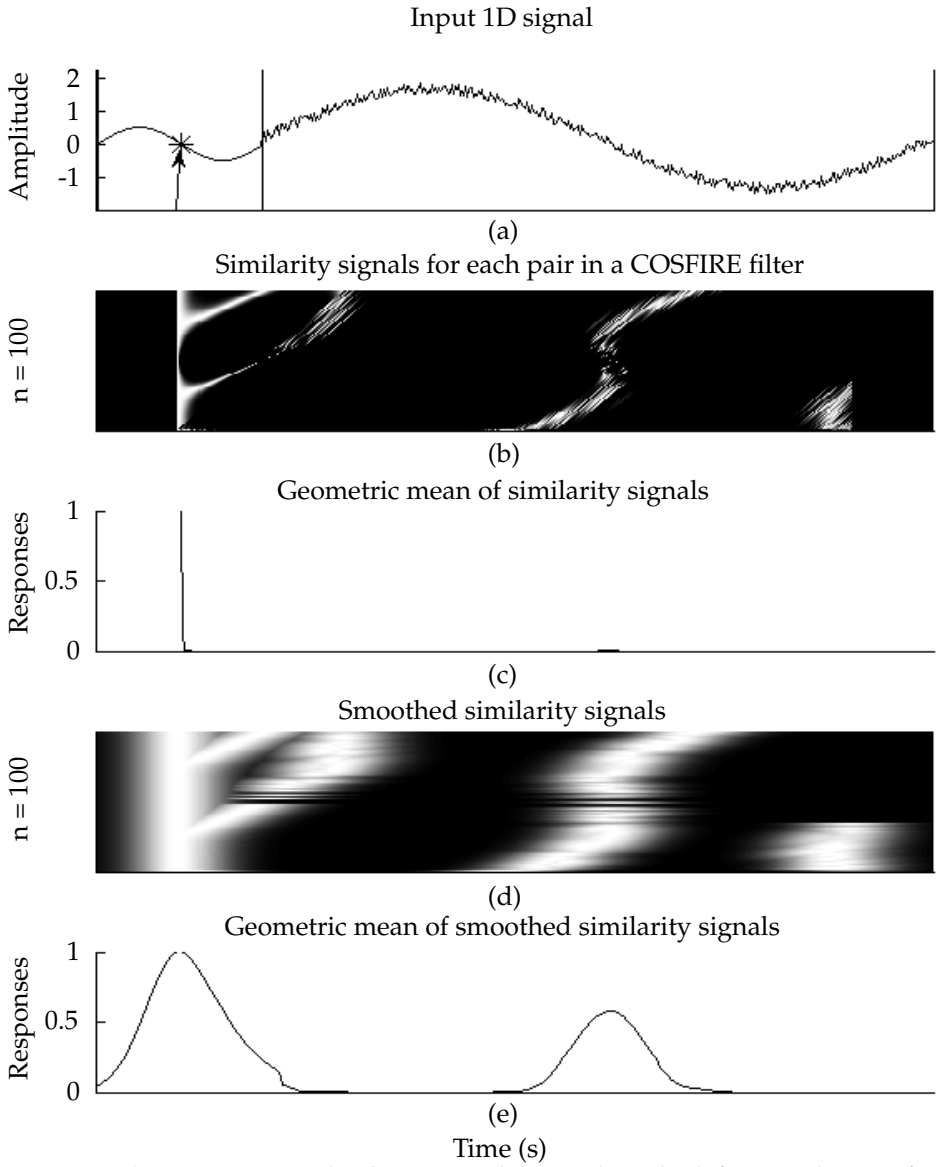

Figure 5.5: a) 1D input signal. The sinusoidal signal on the left is used to configure a COSFIRE filter with parameters $n = 100$, $\delta = 1$, $\sigma_0 = 0.01$ and $\alpha = 0.001$. The arrow indicates its support center. b) The similarity signals between the preferred values of the COSFIRE filter and the values of the input signal. c) The geometric mean of the similarity signals. d) The smoothed similarity signals. e) The COSFIRE filter responses computed as the column-wise geometric from the matrix in (d).

Response

We denote by $R(t)$ the response of a COSFIRE filter (Fig. 5.5c) at time t that is computed as the geometric mean of all the involved smoothed similarity values:

$$R(t) = \left(\prod_{i=1}^{n} S'_i(t) \right)^{\frac{1}{n}} \tag{5.4}$$

Configuration of one COSFIRE filter

In Fig. 5.6a we present an example of the audio signal of the song #22 from our data set, which consists of four motifs. The start and end positions of the vocal pauses are manually annotated with vertical lines. We connect consecutive motifs using linear interpolation in order to fill in the gaps with real numbers and thus avoiding numerical errors.

In this example, the first motif is similar to the third and the second motif to the fourth. All of the four motifs are similar in their first half. This is common in such singing data and it is important that the methods for the identification of melodic repetitions identify them as dissimilar, since the second half is different. For instance, in the example shown in Fig. 5.6a, the first two motifs should be classified as dissimilar. The AND-type character of the COSFIRE filters allows the discrimination of two motifs even when they are similar in their first half. The product of the low similarities in the second half of the motifs and the high similarities of the first halve results in a low COSFIRE response. As a result, the two signals will not be considered similar.

We use the training motif in Fig. 5.6a to construct one COSFIRE filter. In this example we use the fundamental frequency and the time positions of the prototype with $n = 1436$ points, $\sigma_0 = 0.05$ and $\alpha = 0.01$ for the configuration. Then we apply this COSFIRE filter to the entire input signal that consists of four motifs. In Fig. 5.6c we plot the response of the COSFIRE filter for each point of the input signal. The highest response is located in the middle of the training motif with a value of 1. For this and in the remaining parts of the signal, the local maxima are marked with dark spots. The second local maximum response is located in the middle of the third motif. Significantly lower local maxima appear for motif 2 and 4. This is due to the similarity of the motifs only in the first half.

Configuration of multiple COSFIRE filters

As an alternative to the above single filter approach, we split the prototype signal into small parts of equal length and configure a COSFIRE filter by each such part. This is inspired by the hierarchical object recognition approach proposed in

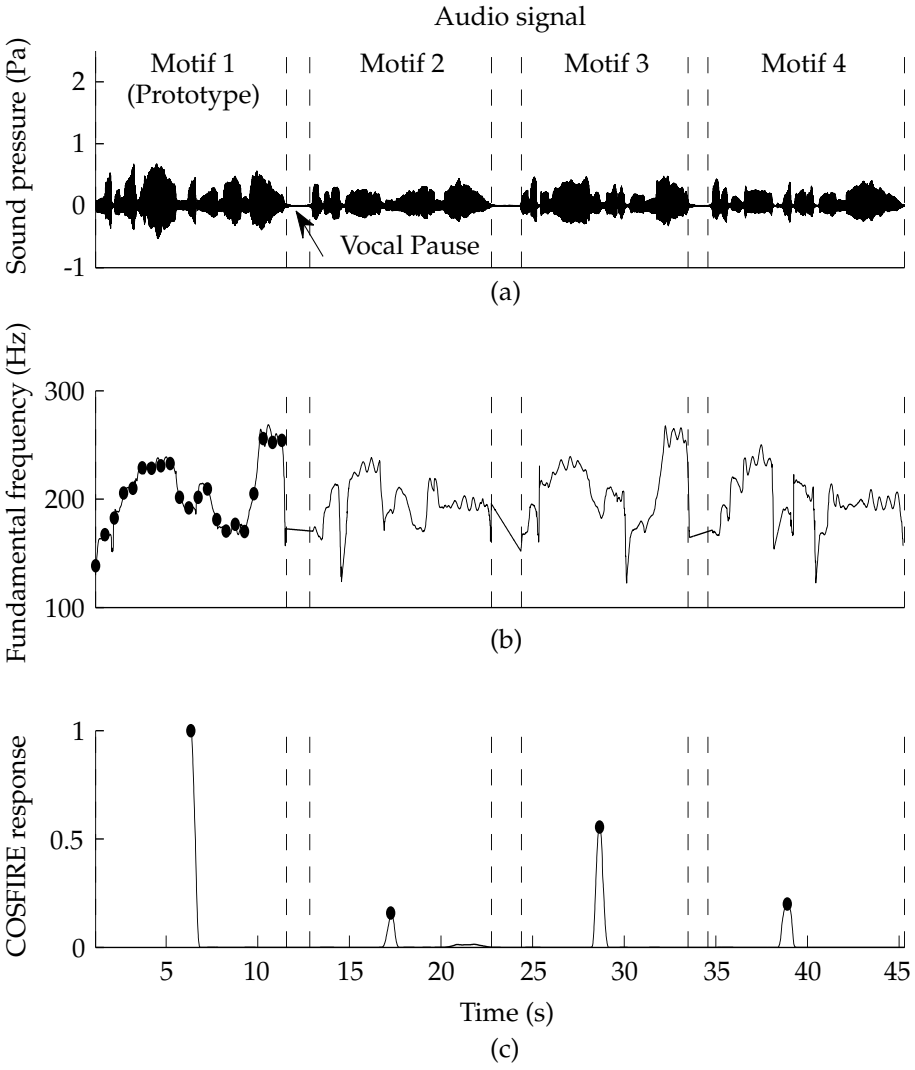

Figure 5.6: a) Audio signal of the song #22 - "Afkoritissa". b) Fundamental frequency of the audio file. Four motifs are manually selected. The first motif is used as a prototype to construct a COSFIRE filter. c) The response signal of the configured COSFIRE filter. The black spots indicate the local maxima points.

(Azzopardi and Petkov 2014). It turns out that the system becomes significantly faster and more accurate when multiple COSFIRE filters with smaller areas of support are used. The advantage of having multiple filters selective for shorter signals is that they allow for more deformation between different parts of the signals, yet

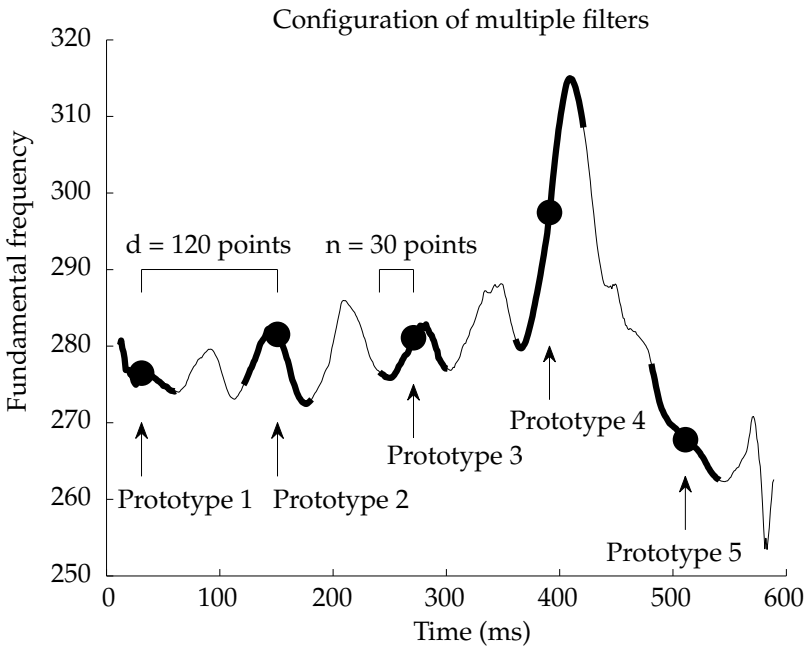

Figure 5.7: Example of the fundamental frequency of a training motif of song #2 - "Tis sousas 2". Five COSFIRE filters are configured with $n = 30$ and distance d between prototype signals is equals to 120 points.

they keep certain rigidity within local parts. For a training motif, we construct a COSFIRE filter every d points, a parameter that defines the distance between the support centers of the concerned filters. In Fig. 5.7 we illustrate an example of a training motif where $(K =)$ 5 COSFIRE filters are configured. The filters are configured with prototypes of width $(n =)$ 30, the intervals between the time points of length ($\delta =$) 1 and the prototype centers separated with a distance of $(d =)$ 120 time points. The thick curve segments show the prototype signals that are used for the configuration of the 5 filters.

In Fig. 5.8a we use the training motif in Fig. 5.6 to construct $(K =)$ 15 COSFIRE filters with $n = 20$, $\delta = 1$ and $d = 90$. The centers of the filters are shown with black spots on the training motif. The response signals of the 15 COSFIRE filters $R_{k=1..K}(t)$ are illustrated in Fig. 5.8b.

The final response is the combination of all the response signals from the five filters. For every time position t we obtain a similarity value that indicates the degree of similarity between the input motif and the training motif. To achieve this combination, it is necessary that all the response signals are aligned to the center of the training motif as shown in Fig. 5.8c. This is done by shifting each response

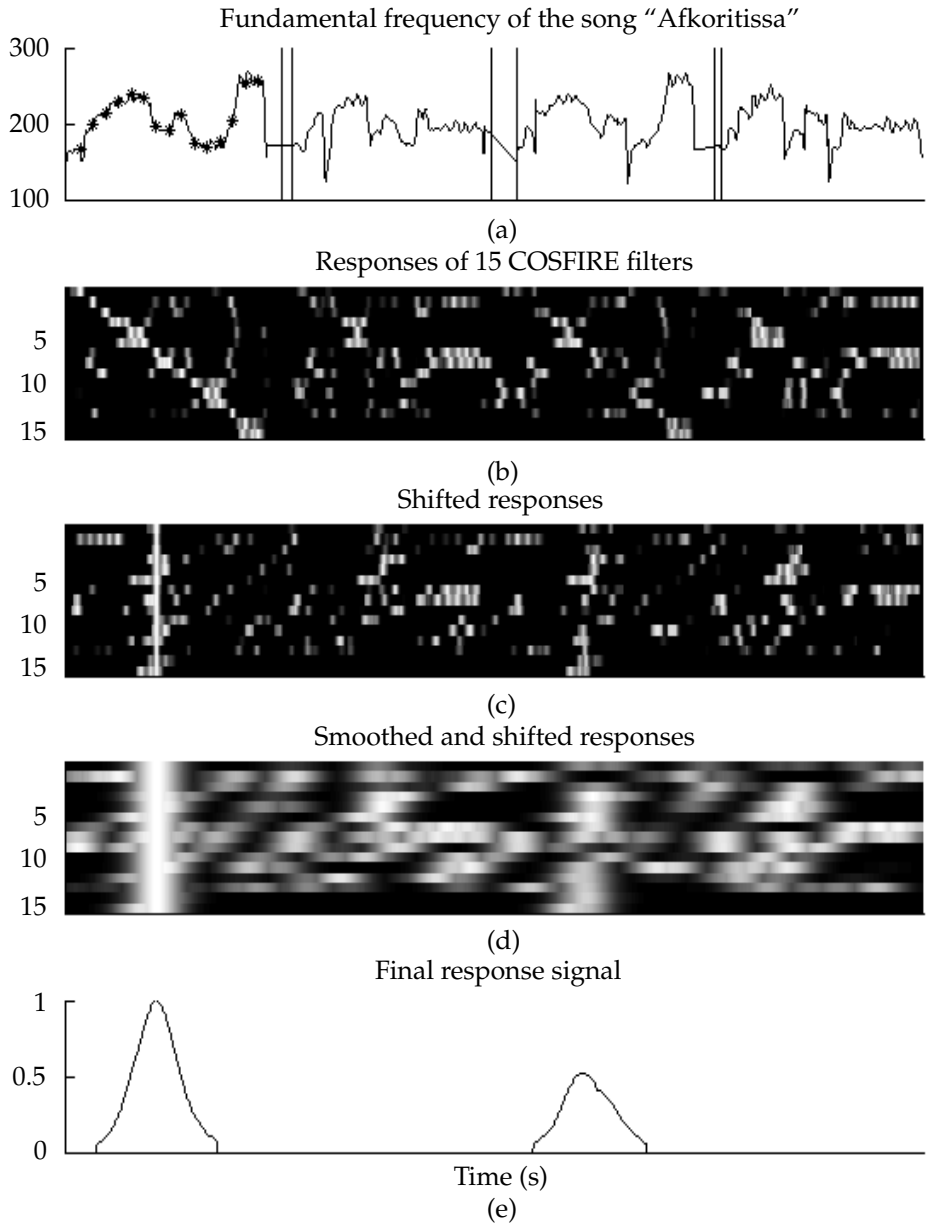

Figure 5.8: (a) Fundamental frequency of the song #22 - "Afkoritissa". The centers of 15 COSFIRE filters are shown with black spots. (b) The responses of every COSFIRE filter. (c) The shifted responses. (d) The blurred shifted responses. (e) The final response.

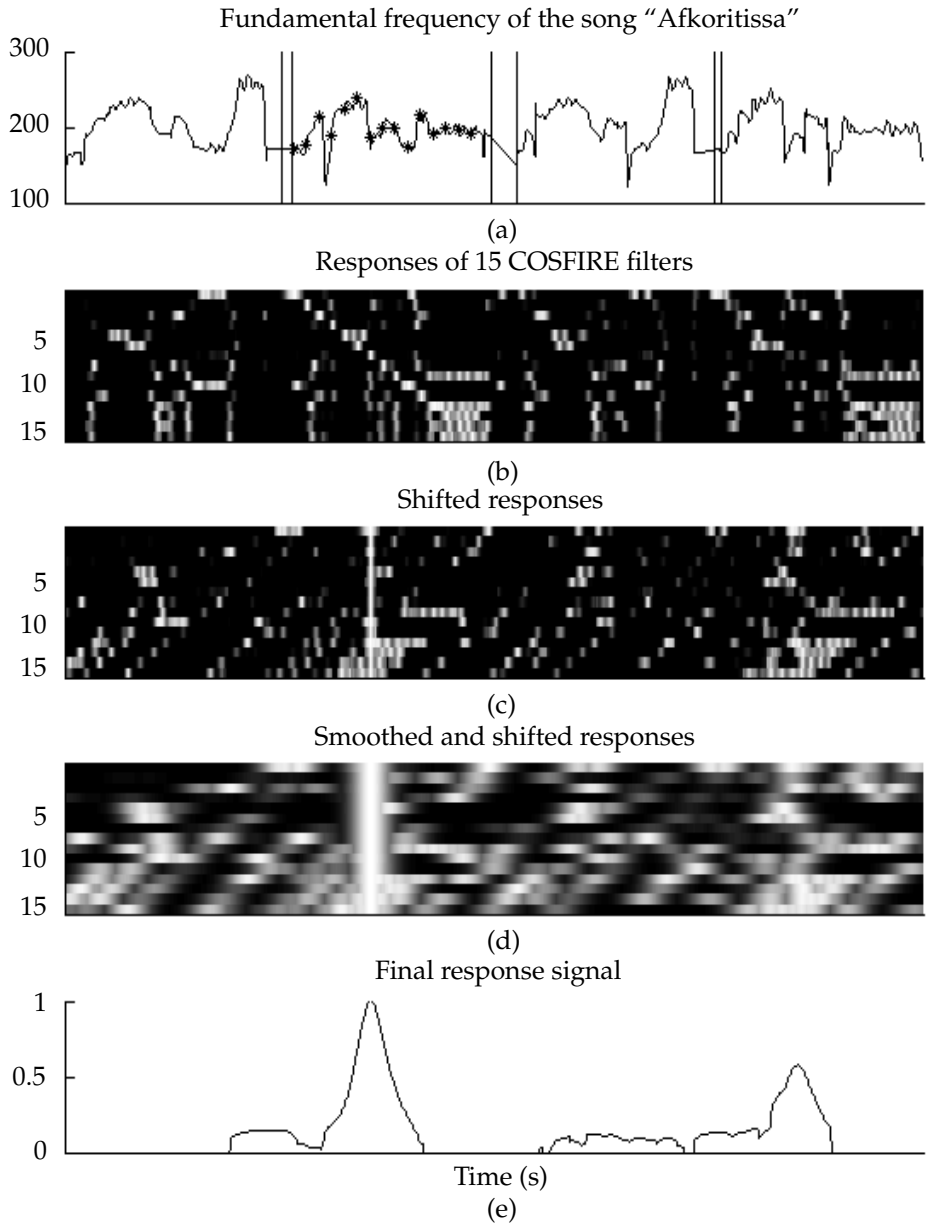

Figure 5.9: (a) Fundamental frequency of the song #22 "Afkoritissa". 15 COSFIRE filters are configured using the fundamental frequency of the second motif. (b) The responses of every COSFIRE filter. (c) The shifted responses. (d) The blurred shifted responses. (e) The final response signal.

signal $R_k(t)$ by the distance from the corresponding prototype k to the center of the training motif. Then, we apply the same smoothing function (with standard deviation $\hat{\sigma}$) explained in Section 5.2.3 to the shifted response signals defined as $R'_k(t)$. The final response is computed as the geometric mean of the smoothed and shifted responses:

$$R(t) = \left(\prod_{k=1}^{n} R'_k(t) \right)^{\frac{1}{n}} \tag{5.5}$$

The smoothed and shifted responses and the final response signals are shown in Fig. 5.8d and 5.8e, respectively. It is shown in this example that, besides motif 1, which was used for configuring this hierarchical COSFIRE filter, we obtain a local maximum in the middle of the third motif. We label motifs 1 and 3 with the letter "A". In Fig. 5.9 we take the second motif as a prototype signal and label it with the letter "B". We follow the same procedure explained above to configure a hierarchical COSFIRE filter and achieve the response signal in Fig. 5.9e. Motifs 2 and 4 are similar and they are represented with the letter "B".

In order to make the process more efficient we apply each low level COSFIRE filter in a specific area of the test motif which we call the response area. First we compute the relative position of the center of the filter in percentage with respect to the normalized duration of the training motif. To find the response area, we multiply this percentage with the duration of the test motif. Then, we apply the COSFIRE filter in an area of three times the value of the n parameter around the relative position in the test signal. Even if the training and the test motifs have a considerable difference in duration but the melody is similar, the approach will detect their similarity.

5.3 Experiments and Results

Here, we demonstrate the effectiveness of the proposed 1D COSFIRE filters in the identification of melodic patterns in monophonic singing folk tunes.

5.3.1 Data

We have created a new data set that consists of 38 audio files of monophonic singing folk tunes of Cyprus with a total duration of 89 minutes (average of 2.3 and standard deviation of 1.2 minutes) and 878 motifs. In all of the songs the performer is a different person. In Table 5.1 we present the data used together with their properties sorted in ascending order of their durations. It is observed that the number of motifs roughly increases linearly with the durations of the songs. The songs were

performed by different persons in different locations and time periods. Of the 38 audio files, 12 were provided by Michalis Terlikkas[2], who is a researcher and experienced musician of the folk music of Cyprus. The remaining 26 songs were collected from the freely available database that is provided by the Intercollege University[3]. The full names of the songs are reported in footnote[4].

The equipment used for the recordings remains unknown. The main characteristic of these songs is that they are monophonic singing folk tunes of Cyprus. These are melodies that were transmitted orally for more than 500 years. Considering the geographical area of Cyprus that is an island in the Eastern Mediterranean sea and the conquerors that influenced the people in several aspects through the years, it is highly probable that cultural elements of those countries were also adapted to the local folk music. Therefore, this data set has a particular value also in the field of ethnomusicology.

5.3.2 Ground truth data

All of the songs in our data set were manually segmented into motifs by the aforementioned researcher of the folk music of Cyprus, Michalis Terlikkas. We used the annotation tool from the WaveLab software[5] and exported the start and end positions of each motif in text format. Then we labelled the first motif of each song with the letter "A". All subsequent motifs that are similar to the first one of each song also take the same label "A". Then we gave the letter "B" to the next unlabelled motif and all subsequent ones similar to it. We repeated this procedure until all the motifs were labeled.

5.3.3 Pre-processing

We apply four pre-processing steps namely error correction, outlier removal, detrending and scaling.

The output of the YIN algorithm may contain two types of errors. The first is the failure of fundamental frequency candidates for non-harmonic sounds. Typically,

[2]http://www.mousalyra.com.cy/en/cv-2/michalis-terlikkas

[3]http://www.cmn.intercollege.ac.cy

[4]S1) Tis sousas 1, S2) Tis sousas 2, S3) Nanourisma 1, S4) Nanourisma 2, S5) Tis sousas 3, S6 Ta mathkia ta giallourika, S7 Sto mnima tou Giorkou, S8 Nanourisma 3, S9 Sto panw mana to xorko, S10 Erotika distixa mavromata, S11 Manes, S12 Tis sousas, S13 Imoun Poulli petameno, S14 Nanourisma 4, S15 I vrisi twn pegeiwtisswn, S16 T ai filipou, S17 Agapw tin je agapa me, S18 Agapisa tin pou karkias, S19 Gia tin kerinia, S20 Kotsini trantafilia, S21 Ise stillos tis karkias, S22 Afkoritissa, S23 Arodafnousa, S24 Nanourisma 5, S25 Makari na rtan oi lampres, S26 Na so tzio armatolos, S27 Karafkiotiki foni, S28 Stile me mana sto nero, S29 Tessera tziai tessera, S30 Psintri vasilitzia mou, S31 Ploumisma fkiolaridon, S32 Armatolos, S33 Ta enia adelfia, S34 Arxi pou kamnw, S35 O Pramateftis, S36 To stolisma, S37 Tis Panagias Tou Kykkou, S38 Nekalima panagias

[5]http://www.steinberg.net/en/products/wavelab/start.html

Table 5.1: Details of the new data set composed of 38 monophonic songs.

Name	Duration (s)	# Motifs	Name	Duration (s)	# Motifs
S1	53.3	14	S20	120.9	20
S2	65	6	S21	121.9	12
S3	67.6	15	S22	123.6	20
S4	67.7	14	S23	125.8	24
S5	70.9	8	S24	129	18
S6	74.9	18	S25	149.7	25
S7	75.2	15	S26	152.7	24
S8	78.2	10	S27	163.2	54
S9	78.7	16	S28	165.1	36
S10	80.6	16	S29	171.8	20
S11	84.6	8	S30	187.2	56
S12	84.6	6	S31	212.1	31
S13	89.5	14	S32	240.8	30
S14	98.6	17	S33	243.5	34
S15	106.6	8	S34	254.4	26
S16	106.6	21	S35	270.1	69
S17	106.6	16	S36	274.5	32
S18	112.3	24	S37	286.4	29
S19	116.6	18	S38	324.6	54

the first local maximum of the output of the ACF represents the fundamental frequency of a periodic signal. For non-periodic signals, such as noise, no fundamental frequency exists and the ACF returns no local maxima. In monophonic songs, non-periodic sounds appear within the vocal pauses. This fact results in the YIN algorithm producing artefacts as candidates for fundamental frequency. In order to correct such errors, we applied a method that was proposed in (Panteli 2011) to remove the wrong values and replaced them with linear interpolation.

The second type of errors of the YIN algorithm are the octave or fifth errors and they appear in situations where the algorithm erroneously outputs the frequency of one of the harmonics of the fundamental frequency. In order to correct such errors, we used the function *FILTERF0.m* from the MAKAM toolbox that is developed by Bozkurt (Bozkurt 2008).

In folk monophonic music, it is common that singers gradually change the pitch while singing. To illustrate an example, in Fig. 5.11a we show the fundamental frequency of the song "O Pramateftis", which has an upward trend. In order to remove the trend in the signal, we subtract from it the best-fit polynomial of order 2. As for outlier removal, we identify and remove those fundamental frequency values

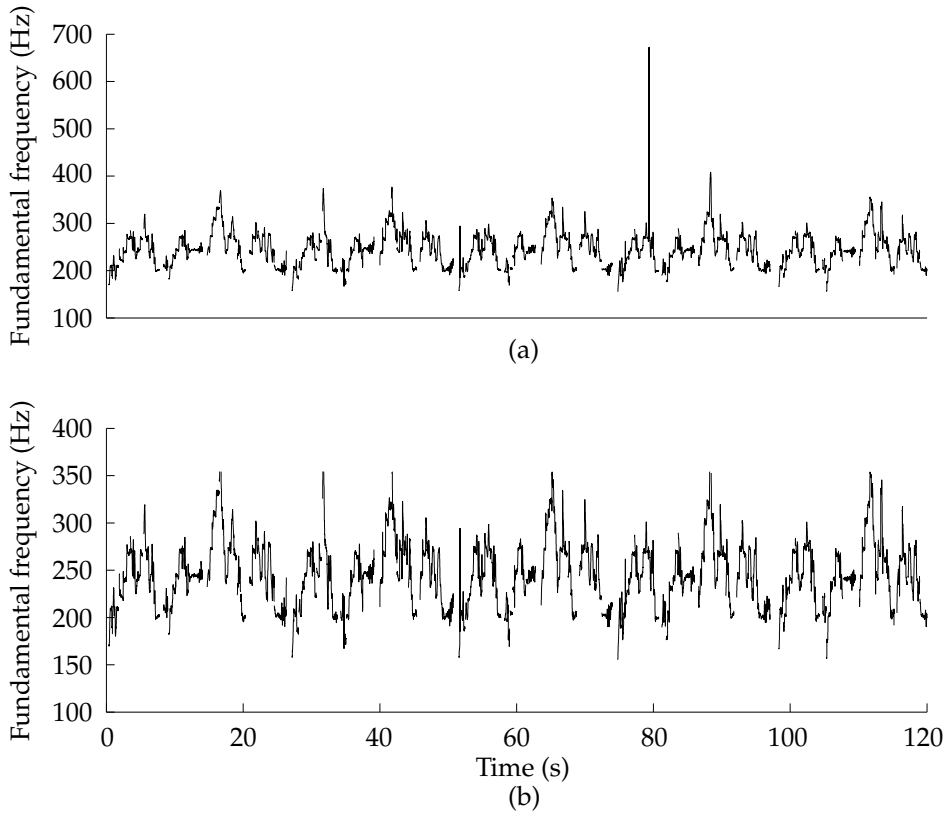

Figure 5.10: (a) The fundamental frequency of the song #20 "Kotsini trantafilia" before and (b) after the outlier removal.

that are three standard deviations away from the mean of the entire signal. Finally, we apply min-max scaling such that the data points become in the range [0,1]. In this way, the method becomes transposition invariant. The preprocessed signal is shown in Fig. 5.11b.

5.3.4 Tuning of Parameters

We split the data set into an evaluation and a test set. The evaluation set consists of the first 19 songs that are listed in Table 5.1, which we use to fine tune the parameters of the proposed COSFIRE-based approach that is characterized by six parameters; width of prototype sub-motifs n, distance between prototype sub-motifs d, frequency tolerance parameters σ_0 and α, as well as temporal parameters σ' and $\hat{\sigma}$. In practice, we perform a grid search over the sets of parameter values: $n = \{30, 50, 70\}$, $d = \{30, 50, 70\}$, $\delta = \{1, 10, 20\}$, $\sigma_0 = \{0.001, 0.005, 0.01\}$,

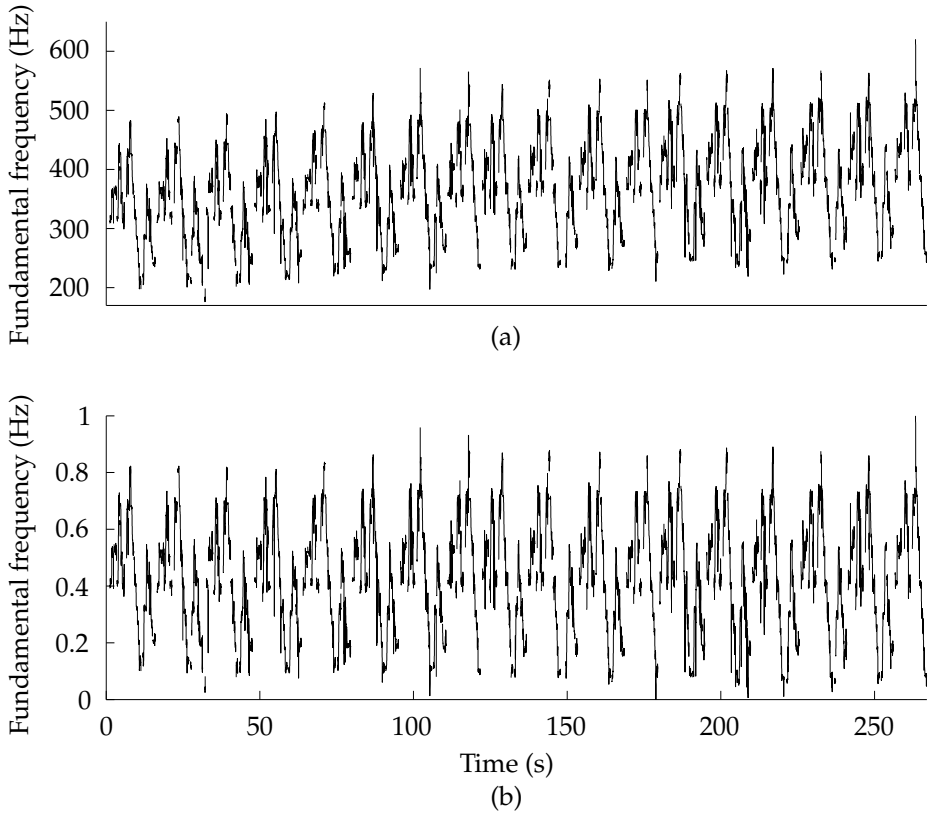

Figure 5.11: (a) The fundamental frequency of the song #35 "O Pramateftis" before and (b) after detrending.

$\alpha = \{0.001, 0.005, 0.01\}$, $\sigma' = \{0.001, 0.005, 0.01\}$ and $\hat{\sigma} = \{0.1, 0.5, 1\}$.

The output of the proposed algorithm to a given audio signal is a string of letters and is obtained by the method explained in Section 5.2.1. Its length is equal to the number of motifs in the given audio signal. Then, we compute the Hamming distance between the resulting string and the string given in the ground truth and divide with the length of the strings. The set of parameters $n = 30$, $d = 50$, $\delta = 1$, $\sigma_0 = 0.001$, $\alpha = 0.01$, $\sigma' = 0.01$ and $\hat{\sigma} = 0.5$ achieved the minimum mean Hamming distance over the validation set.

5.3.5 Comparison to other methods

We compare our results with other popular methods in signal processing namely cross correlation, Dynamic Time Warping (DTW) and Symbolic Aggregate approXi-

mation (SAX) (Keogh et al. 2006). The DTW is widely used in MIR applications and in similar problems such as the identification of melodic repetitions.

Cross Correlation

It slides one signal on the other and computes the dot product at each position. The maximum value is used as a measure of similarity between the two signals. We tune the threshold parameter for the dissimilarity classification. We compute the Precision and the Recall for a set of threshold values and we choose the value that returns the maximum F_1 score.

Dynamic Time Warping (DTW)

The DTW method has been used in a variety of applications in signal processing, particularly in MIR and in speech recognition (Peiszer 2007, Shiu et al. 2005, Jehan 2005, Muda et al. 2010). The main benefit of this method is that it allows for temporal tolerance when comparing two time signals with different durations. Other methods, such as cross correlation, are less robust in this respect. The DTW calculates the optimal path between two given time signals and it returns a value between 0 and 1. A value of zero indicates identical signals. This method has the threshold parameter to tune for the dissimilarity classification.

Symbolic Aggregate Approximation (SAX)

The SAX method computes a symbolic representation of a 1D signal by assuming that the data points are normally distributed. It uses two parameters, the size of a vocabulary and the length of the resulting symbolic string. We standardize the given signal to have a zero mean and a standard deviation of 1. For the technical details on this method we refer the reader to (Keogh et al. 2006).

For the three above mentioned methods we use a similar procedure that is explained in Section 5.2.1 to obtain a string representation. For a given audio signal, we take the first motif and use the concerned method (cross correlation, DTW and SAX) to obtain a similarity value for the remaining motifs in the same signal. We give the letter "A" to the first motif and to the motifs for which a similarity value above a certain threshold is achieved. Then, we take the next motif that has not yet been labeled and compare it with the remaining unlabelled motifs, and give the letter "B" to the ones that evoke a response greater than the threshold. We repeat this process until all motifs are labeled. For the SAX method we use the minimum (lower-bounding) distance[6] suggested in (Keogh et al. 2006) to compute the similarity between the symbolic representations of the motifs. Moreover, we use the

[6] In practice we use the Matlab implementation provided in http://www.cs.ucr.edu/∼eamonn/SAX.htm

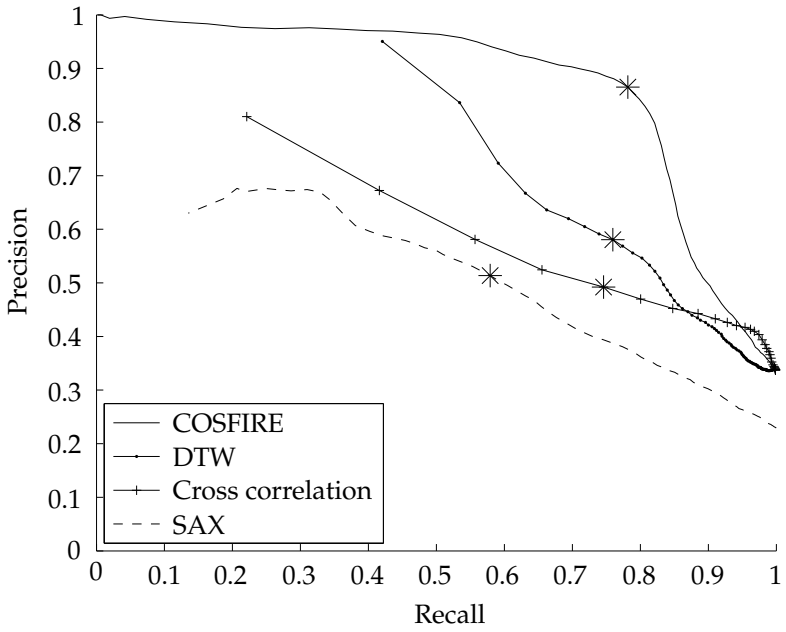

Figure 5.12: The precision and recall achieved by the proposed COSFIRE approach in comparison to cross correlation, DTW and SAX, as a function of the involved threshold that is common to all methods. The markers indicate the precision and recall at the maximum f-score.

validation set to fine-tune the involved two parameters by a grid search. The parameters that achieve the minimum mean Hamming distance are a vocabulary of 8 letters and a length of 16 letters in the resulting symbolic string. The other two methods have no specific parameters to tune other than the threshold parameter.

5.3.6 Results

The test set consists of the last 19 songs that are listed in Table 5.1. For the COSFIRE approach that we propose and for the SAX method we use the parameter values that returned the best results from the grid search on the validation set. For the other two methods we tune the threshold that is applied for the dissimilarity classification. We use a different threshold for each method that returns the maximum F_1 score over a set of threshold values.

We present the results of the four methods (COSFIRE, DTW, cross correlation and SAX) that we applied to our data set in terms of Precision P=TP/(TP+FP) and Recall R=TP/(TP+FN) and the harmonic mean also known as F_1 score 2PR/(P+R).

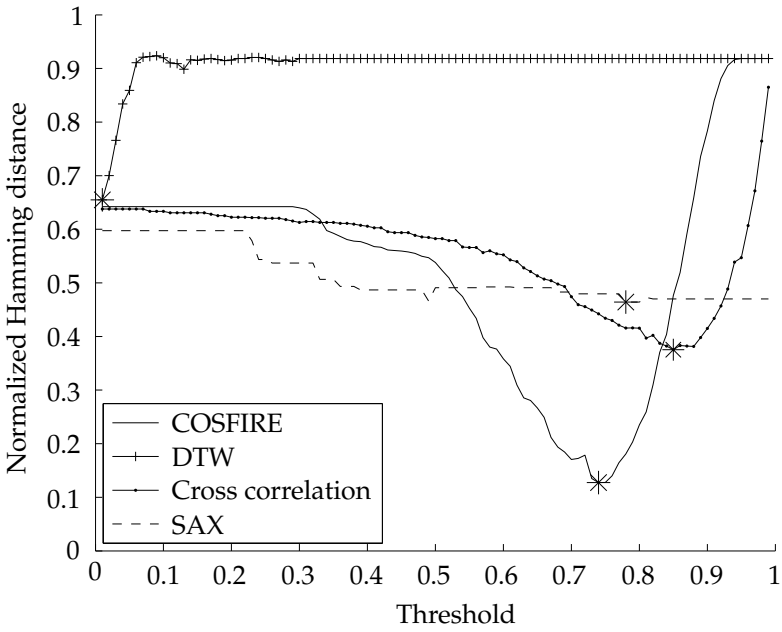

Figure 5.13: The sum of normalized Hamming distances achieved by the proposed COSFIRE approach in comparison to cross correlation, DTW and SAX, as a function of the involved threshold that is common to all methods. The markers indicate the minimum point of each plot.

For the symbolic representation matching, we calculate the normalized Hamming distance between the strings of the ground truth and the output of each of the four methods. Then, we obtain the mean Hamming distance over the entire test set for various values of the threshold th.

The TP, FP and FN are the abbreviations for True Positives, False Positives and False Negatives, respectively. A classification is called TP if a pair of motifs that are compared are annotated as similar in the ground truth and the method classified them as similar. A FP is the wrong classification of the system that two motifs are similar, while a FN is the wrong classification that two motifs are dissimilar.

The classification of a pair into similar or dissimilar has been done using a threshold on the output of the respective method. We have computed results for 101 values of the output threshold in the range 0 and 1 in intervals of 0.01. These results are presented in a precision and recall plot in Fig. 6.8. In Fig. 5.13 we plot the mean of Hamming distances over the test set for the four methods as a function of the threshold th.

The proposed COSFIRE approach outperforms the other methods. The har-

monic mean of the precision and recall reaches a maximum value of 0.83 at a recall R of 0.80 and a precision P of 0.86. The precision and the recall of the SAX method at its maximum harmonic mean of 0.46 is 0.41 and 0.52, respectively. The harmonic mean of the precision and recall of the DTW reaches its maximum $F_1 = 0.66$ at a (R = 0.74, P = 0.59). The maximum harmonic mean of the cross correlation is achieved with $F_1 = 0.60$ at a (R = 0.75, P = 0.5).

For the symbolic representation, COSFIRE achieves a minimum mean of Hamming distances of 0.13 with a threshold th = 0.74. On the other hand, the minimum mean of Hamming distances of the cross correlation, DTW and SAX are 0.66 (th = 0.01), 0.38 (th = 0.85) and 0.39 (th = 0.46), respectively.

Additionally, we calculated the processing time that each method requires to process every song. The execution time of the methods COSFIRE, SAX and cross correlation to compute 38 songs were 51.5, 0.9 and 0.1 minutes respectively, while that of the DTW was 729.6 minutes.

For each of the four methods and for each song, we present in Table 5.3.6 the F_1 score, the execution time and the normalized Hamming distance for the threshold that contributes to the minimum F_1 score and mean of Hamming distances on the test set. The execution time is calculated with sequential implementation of the methods running on a personal computer with a 1.7 GHz processor and 8 GB of RAM.

5.4 Discussion

The COSFIRE approach performs substantially better than the cross correlation, DTW and SAX methods. Most of the existing methods for the identification of motifs make use of the DTW in their methodology. Therefore, we believe that replacing DTW by the proposed COSFIRE approach may yield significant improvement in several applications. While the SAX and the cross correlation methods are more efficient than COSFIRE, they are much less effective possibly due to the insufficient robustness to temporal tolerance.

In contrast to the QBH method that has been a state-of-the-art for melodic similarity tasks, COSFIRE filters compare pitch tracks that are derived directly from the audio signal. In other words, the note segmentation and the database with the MIDI transcriptions are avoided. Similarly, wavelets have been effective in applications that are using the pitch tracks. Unlike COSFIRE, wavelets are linear functions and are not intrinsically robust to temporal tolerance. They are typically used to extract features that can be used in a classification model. This is in contrast to the proposed COSFIRE filtering approach that can be used directly as a similarity function without involving classification models.

Several data sets for folk music analysis are used in the literature, while not all

Table 5.2: The excecution time (ET) and the normalized Hamming distances (NHD) of each song for the threshold that contributes to the minimum mean of hamming distances for the four methods.

	COSFIRE			DTW			SAX			Cross Correlation		
F_1 score	th = 0.66			th = 0.91			th = 0.71			th = 0.95		
NHD	th = 0.74			th = 0.85			th = 0.78			th = 0.01		
Song #	ET (min)	NHD	F_1 score	ET (min)	NHD	F_1 score	ET (min)	NHD	F_1 score	ET (min)	NHD	F_1 score
20	0.41	0.00	0.69	17.40	0.50	0.72	0.24	0.75	0.52	0.39	0.75	0.41
21	0.25	0.00	1.00	15.09	0.25	0.78	0.08	1.00	0.41	0.08	0.67	0.71
22	0.39	0.25	1.00	12.18	0.15	0.88	0.18	0.60	0.87	0.20	0.75	0.58
23	0.57	0.04	1.00	16.15	0.21	0.61	0.30	0.50	0.26	0.24	0.75	0.73
24	0.44	0.28	0.67	16.87	0.17	0.59	0.16	0.11	0.22	0.16	0.83	0.54
25	0.70	0.00	1.00	22.72	0.08	0.95	0.30	0.00	0.85	0.22	0.36	0.82
26	0.64	0.00	1.00	20.71	0.71	0.96	0.30	0.00	0.33	0.24	0.50	0.87
27	1.54	0.04	0.97	23.97	0.56	0.95	1.17	0.59	0.63	0.81	0.67	0.71
28	1.10	0.31	0.90	23.62	0.58	0.83	0.55	0.42	0.90	0.45	0.42	0.53
29	0.63	0.05	1.00	25.93	0.05	0.97	0.20	0.00	0.63	0.20	0.50	0.89
30	1.92	0.02	0.99	30.86	0.68	0.89	1.51	0.79	0.77	1.12	0.82	0.41
31	1.20	0.03	0.76	37.02	0.71	0.60	0.44	0.65	0.57	0.41	0.68	0.60
32	1.26	0.00	0.85	43.52	0.07	0.93	0.46	0.33	1.00	0.41	0.50	0.87
33	1.41	0.47	0.87	47.24	0.15	0.80	0.54	0.97	0.40	0.55	0.50	0.65
34	1.09	0.08	0.98	49.62	0.12	0.66	0.38	0.08	0.82	0.33	0.50	0.65
35	2.22	0.02	0.59	61.73	0.33	0.68	1.18	0.00	0.63	0.86	0.80	0.74
36	1.38	0.22	0.86	56.77	0.66	0.88	0.52	0.66	0.72	0.48	0.84	0.68
37	2.66	0.55	0.76	62.88	0.67	0.60	1.47	0.62	0.79	1.16	0.84	0.66
38	3.00	0.07	0.96	79.15	0.50	0.78	1.37	0.76	0.65	1.37	0.76	0.59
Mean	1.20	0.13	0.89	34.92	0.38	0.79	0.60	0.46	0.63	0.51	0.65	0.66
Standard deviation	0.79	0.17	0.13	19.68	0.25	0.14	0.48	0.34	0.23	0.38	0.16	0.14

of them are publicly available. We have tried to apply our method in the MTC-ANN-2.0 data set from the Meertens Tune Collections[7] that is used in (Rodrıguez-López and Volk n.d., Janssen et al. 2015, van Kranenburg and Tzanetakis 2010) for tasks such as melodic segmentation and motif identification. We have encountered some difficulties related to the consistency of the data sets that are used in (Janssen et al. 2015) and the publicly available data set, in order to compare our results. In (Janssen et al. 2015), the experiments are done using a subset of the MTC-ANN-2 data set of 16 tune families while the published data set contains 28. The data are stored in kern format and thus the music information is encoded with the MIDI protocol. In addition to this, the melodic contour of some of the annotated motifs that are used for training is significantly different than their similar ones that are used for testing. Therefore, it is possible that the annotators used other features than the melody for ranking the similarity of the motifs, such as the rhythmic pattern. For the above mentioned reasons, COSFIRE does not seem to be an appropriate method for that data set. In (Ross et al. 2012) the authors use data from a personal collection and the annotation of the motifs was done manually.

In Fig. 5.14, we present the distribution of the fundamental frequency of the first motif of the song #1 "Tis Sousas 1". The values are concentrated around four frequencies, the ones that represent the most common notes that were performed during the analyzed song. The theoretical frequencies of these notes are C#4 (277.2Hz), D#4 (311.1Hz) and F4 (349.2Hz), which are indicated with dashed vertical lines in Fig. 5.14. This example demonstrates that the fundamental frequency of a song does not follow a normal distribution. In this respect, the SAX method may not be suitable for such signals because it assumes an underlying normal distribution.

The benefit of converting a song into a symbolic representation is twofold. First, it permits data compression and thus allows faster comparison between two musical signals. The symbolic representation is a shorter sequence of musical events that can be used as a higher level feature for classification. Second, from a musicological point of view, the sequence of the repeating motifs is an important and interesting feature in itself. It is very common in musicology to split a song into musical events and report their sequence and the frequency of their appearance. The automatic identification of such events allows the analysis of large data sets in relatively short period of time.

A system that automatically analyzes folk music is very important since there are no written musical scores for musicological analysis. In folk music, musicologists mark manually every musical event for their research. Similar to other applications, manual annotation may result in inaccuracies due to fatigue and is certainly much slower than automated methods.

We present a novel filter based approach for the identification of motifs and a

[7] http://www.liederenbank.nl/mtc/

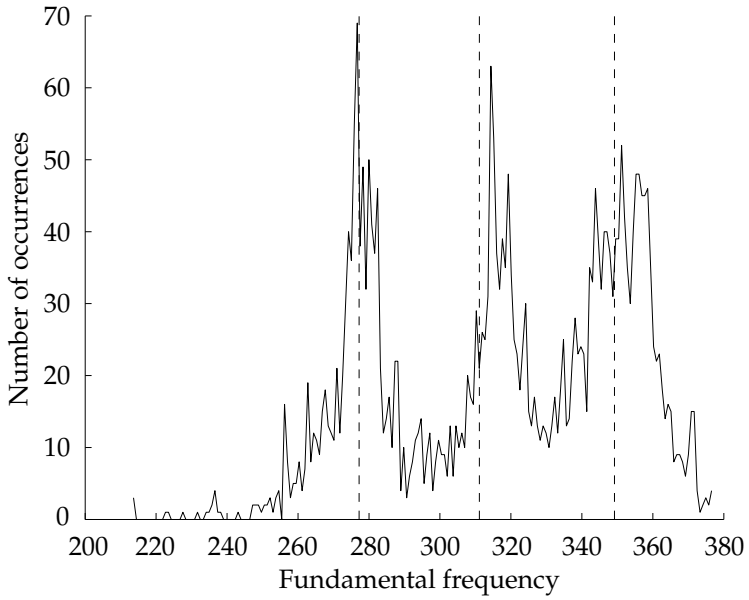

Figure 5.14: The distribution of the fundamental frequency of the first motif of the song #1 "Tis Sousas 1". Dashed vertical lines show the frequencies of the notes C#4, D#4 and F4.

method for representing a monophonic song into a symbolic string. We have shown that the proposed COSFIRE filters allow temporal tolerance but are more effective and more efficient than DTW. This is mainly attributable to the hierarchical structure that COSFIRE uses. DTW gives the same temporal tolerance to every part of the signal irrespective of its position. As demonstrated in our experiments, COSFIRE filters can be configured to take input from other COSFIRE filters that are selective for smaller parts of the signal. This arrangement provides the possibility to have low tolerance in local parts and higher global tolerance between the involved parts.

The COSFIRE filters are trainable and easy to implement. Even though the data set used to apply our method refers to ethnomusicological interest, the proposed COSFIRE filters can be applied in any application that involves 1D signals, such as the identification of k-complexes and spindles in EEG signals (Camilleri et al. 2014).

The computation of a COSFIRE filter response is parallelizable as it relies on independent operations defined in the concerned set of tuples. While this was beyond the scope of this work, we speculate that such parallelization would largely improve the efficiency of the algorithm by a factor that corresponds to the number of tuples.

In this work, we are not dealing with automatic segmentation in motif and note level. Moreover, our data set is a simplistic representation of the general music that

is available as data since we applied our method in monophonic songs. Therefore, the major contribution of the COSFIRE filters that are proposed in this chapter is their use as a distance metric between two sequences. For future work we aim to use the nested COSFIRE filters in order to develop an automated system for motif segmentation. Additionally, we will explore the potential use of other algorithms for polyphonic melody extraction to expand the generalization ability of our method.

5.5 Conclusion

We propose an effective trainable filter approach for identifying motifs in 1D signals, with application to acoustic signals. The transformation of a song into a symbolic representation leads to data reduction and the identification of the most important melodies in a song.

By means of experiments on a new data set of monophonic folk songs, we demonstrated that the proposed method outperforms the existing cross correlation, SAX and DTW methods with a time complexity that is higher than that of the former two methods but substantially lower than that of DTW.

Published as:

A. Neocleous, G. Azzopardi, C. N. Schizas and Nicolai Petkov, "Filter-based approach for ornamentation detection and recognition in singing folk music' *Computer Analysis of Images and Patterns: 16th International Conference (CAIP), Lecture Notes in Computer Science*, Vol. 9256, pp. 558-569, 2015.

Chapter 6

Ornamentation detection using COSFIRE filters

Abstract

Ornamentations in music play a significant role for the emotion which a performer or a composer aims to create. The automated identification of ornamentations enhances the understanding of music, which can be used as a feature for tasks such as performer identification or mood classification. Existing methods rely on a pre-processing step that performs note segmentation. We propose an alternative method by adapting the existing two-dimensional COSFIRE filter approach to one-dimension (1D) for the automatic identification of ornamentations in monophonic folk songs. We construct a set of 1D COSFIRE filters that are selective for the 12 notes of the Western music theory. The response of a 1D COSFIRE filter is computed as the geometric mean of the differences between the fundamental frequency values in a local neighbourhood and the preferred values at the corresponding positions. We apply the proposed 1D COSFIRE filters to the pitch tracks of a song at every position along the entire signal, which in turn give response values in the range [0,1]. The 1D COSFIRE filters that we propose are effective to recognize meaningful musical information which can be transformed into symbolic representations and used for further analysis. We demonstrate the effectiveness of the proposed methodology in a new data set that we introduce, which comprises five monophonic Cypriot folk tunes consisting of 428 ornamentations. The proposed method is effective for the detection and recognition of ornamentations in singing folk music.

6.1 Introduction

A common technique for expressing emotions in music performance is the addition of short notes to the main melody. These notes are called ornamentations and can be arbitrarily or systematically inserted. A glissando, also known as vibrato, for instance, is a rapid alteration of a series of consecutive notes. It is one of the most frequent ornamentations. Several other ornamentations are also used such as amplitude variation called tremolo and stretching or shortening the duration of notes, among others. In Fig. 6.1a we present the audio signal of the song Syrinx

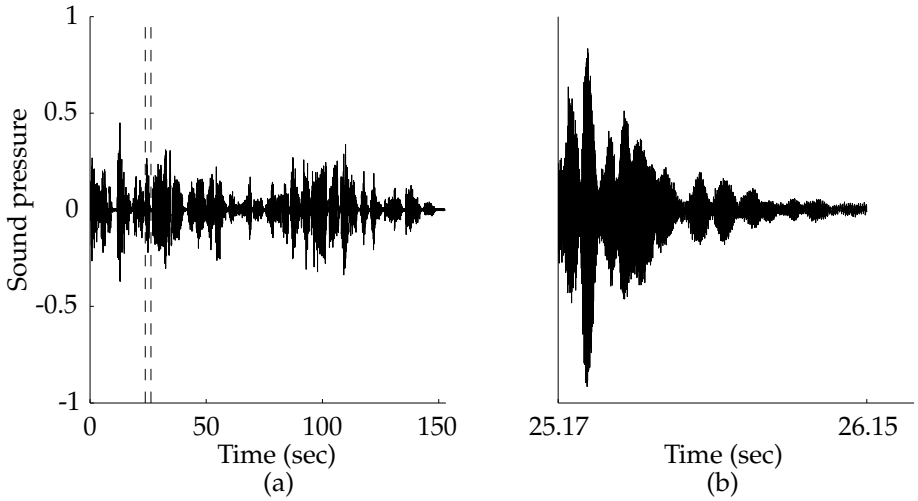

Figure 6.1: (a) The sound pressure signal of the song Syrinx for solo flute by Claude Debussy. (b) Tremolo ornamentation that is characterized by amplitude modulation.

for solo flute by Claude Debussy. The vertical lines indicate a note with a tremolo ornamentation, which is shown enlarged in Fig. 1b.

The importance of ornamentations in music has been researched and described by music theorists (Taylor 1991). They are related to the feeling which a performer or a composer aims to create. In the field of music information retrieval (MIR), it has been shown that musicians have a unique way to perform their ornamentations and that it is a distinctive feature for performer identification (Ramirez et al. 2007, Ramirez et al. 2008, Ramirez and Maestre 2009).

The majority of these related studies are focused on applications in Western music and they mainly analyse ornamentations of musical instruments rather than singing voice (Boenn 2007, Casey and Crawford 2004, Gainza and Coyle 2007). Furthermore, the methodology of these studies require note segmentation. This creates additional computational difficulties since note segmentation is still a challenging problem in MIR.

In (Köküer et al. 2014) the authors initially attempt to segment notes in traditional flute performances. Then they explore knowledge about ornamentations within a segmented note. They propose to separate ornamentations into two categories namely single-note and multi-note, which in turn are composed of two and three-sub categories, respectively.

In this work we are interested in ornamentation detection and recognition of folk music of the Eastern countries. It is generally harder to process Eastern folk songs than Western music from a signal processing point of view. Since folk music is frequently recorded in an environment such as a public place, which may cause low

quality recordings covered with background noise. Also, the singers in folk music are usually not professionals and therefore they may sing out of tune, forgetting melodies and others. One major difference between Eastern and Western music is the frequency distance between consecutive notes. In Western music this difference is strictly logarithmically equal into twelve notes per octave. In Eastern music and especially in Makam this distance between notes is not equally distributed. An automated system that will be able to capture ornamentations from an audio signal will help to accumulate additional knowledge of non-Western folk music, such as Makam in Turkish and Arabic music.

We propose a novel filter-based algorithm for the automatic identification of ornamentations in singing folk music of Cyprus. It is adapted from the two-dimensional COSFIRE approach (Azzopardi and Petkov 2013, Azzopardi and Petkov 2014), which has been demonstrated to be effective in object localization and recognition in images.

This chapter is organized as follows. First we introduce the proposed type of ornamentations in Section 6.2. In Section 6.3 we describe the proposed methodology and demonstrate its effectiveness in Section 6.4. Finally, we discuss certain aspects of the approach that we propose and draw conclusions in Section 6.5.

6.2 Types of ornamentations

Ornamentations in music have been precisely defined mainly in Western music since the 17th century with extensive use in the Baroque period. Since then, the composers have been annotating their desirable ornamentations in the so-called musical score. A musicological study can gather significant information from such transcriptions such as note frequency and duration, rhythm, tempo and others.

In folk music usually the composer is not known, hence there is no written score or any similar information about a song. The music is transmitted orally and mutates over time. Therefore, significant information about ornamentations is stored in the available recordings.

In this chapter, we make an attempt to create meaningful categories that describe the type of each ornament. We adopt the terms “single-note” and “multi-note” ornamentations from (Köküer et al. 2014) and we introduce some additional sub-categories based on the music theory. We propose the following sub-categories within the single-note category: linear positive, linear negative, glissando positive and glissando negative (Fig. 6.2a). The major feature of this category is that only one alteration of a small note is done. A multi-note consists of alterations of more than one note and comprises two sub-categories: the vibrato positive and vibrato negative (Fig. 6.2b).

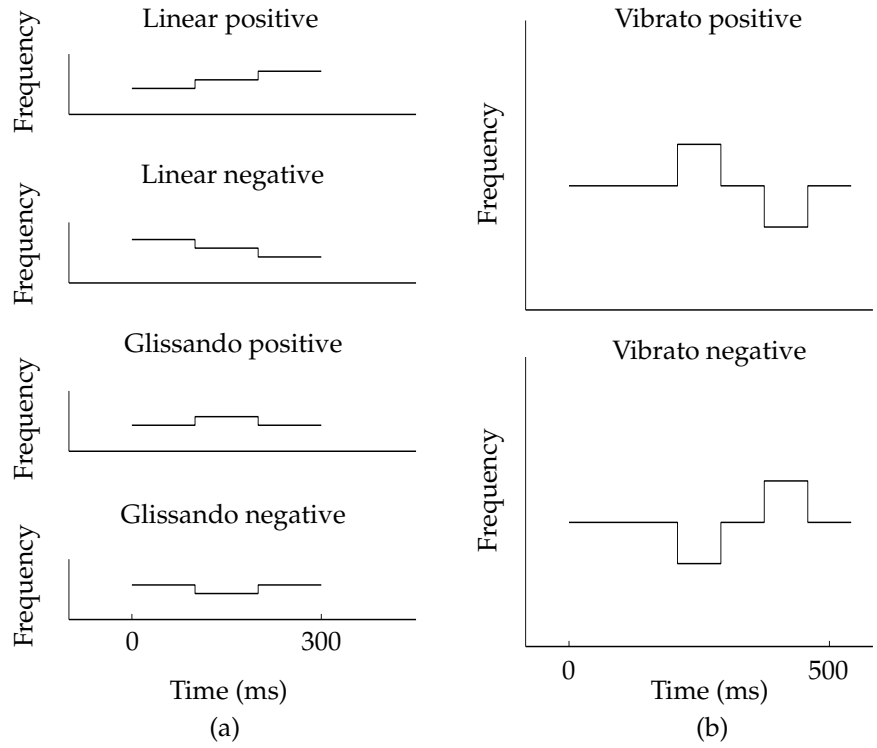

Figure 6.2: Ornamentations which belong to (a) single-note and (b) multi-note subcategories. The frequencies can take any values in the range of singing voice.

6.3 Methods

In this section we present the three main steps of the proposed methodology: feature extraction, configuration and application of the proposed 1D COSFIRE filters for ornamentation detection, followed by detection and recognition. In Fig. 6.3 we present the main steps of our methodology.

6.3.1 Feature extraction

Initially we segment the audio signal in overlapping frames of 33ms duration and 3ms overlap. Then we use the YIN algorithm to extract the fundamental frequency for each audio frame (de Cheveigné and Kawahara 2002). This method belongs to the time-domain based algorithms. First, the Autocorrelation function (ACF) is computed in each frame. Some of the peaks in the output of the ACF represent multiples of the period of the input signal. The repeating pattern of the audio signal is identified by choosing the highest non-zero-lag peaks of the ACF output, and pro-

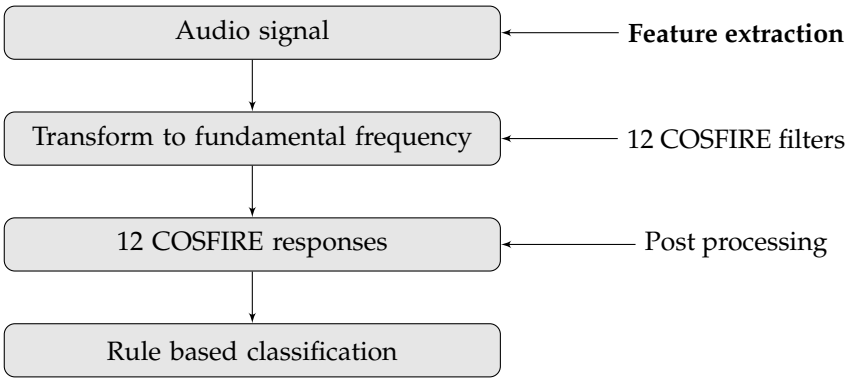

Figure 6.3: The main steps of our proposed method. The audio signal is converted in to its fundamental frequency (pitch track). 12 COSFIRE filters are configured using the frequencies of the 12 western notes. The COSFIRE filters are applied to the pitch track, returning 12 responses, one for each COSFIRE filter. The responses are then binarized and post processed. Rule-based approach is used to identify and classify ornamentations.

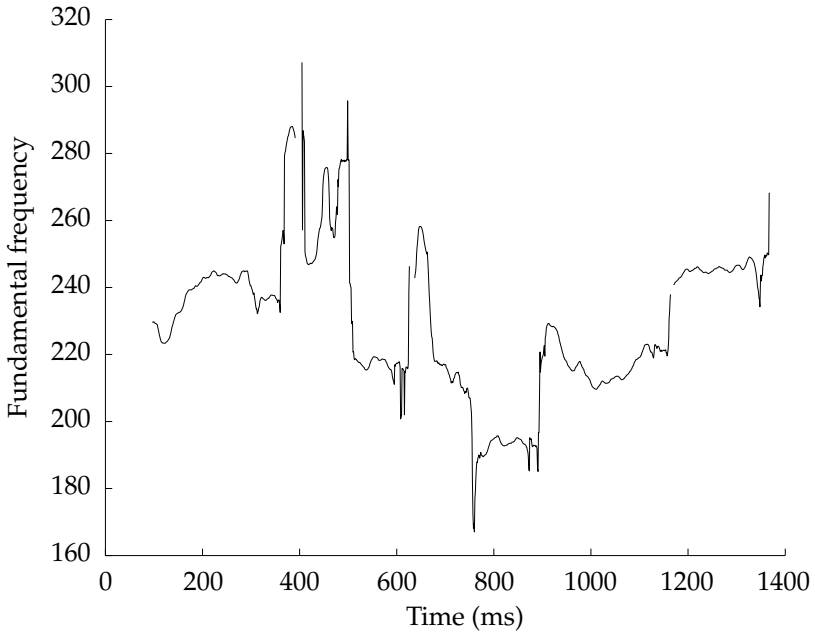

Figure 6.4: Fundamental frequency as a function of time is used as input to our system.

cess them with a number of modifications to return a candidate for the fundamental frequency. We refer to the output of the YIN algorithm as the "pitch track". In Fig. 6.4 we show the pitch track extracted from one of the songs in our data set. The discontinuities in the pitch track are caused by the inharmonicity of the signal at vocal pauses or other non harmonic sounds such as consonants.

We apply a post-processing step to the pitch track in order to correct some errors of YIN. The most frequent error is the so called "octave error". The algorithm erroneously chooses a candidate for a fundamental frequency that has a value in the higher or lower octave. More information about this post-processing method can be found in (Panteli 2011).

6.3.2 Configuration of a COSFIRE filter

Originally COSFIRE filters were proposed as trainable filters for computer vision applications. Here we adopt that idea to 1D pitch tracks. A 1D COSFIRE filter uses as input the frequency values at certain positions around a specific point in time of an audio signal. The preferred frequency values and positions for which the resulting filter achieves a maximum value of 1 are determined in manual configuration process. In theory, a note has a single frequency that is constant over time. In practice however, the frequencies of a singing note vary slightly over time as illustrated by an example in Fig. 6.6a and the shape of the frequency signal is different every time the note is performed. Since one note has a specific duration and the theoretical fundamental frequency of a note is constant over time, we consider the same frequency value for a set of n points and we call this vector a prototype. The parameter n is set experimentally. We choose to configure prototypes with shorter durations as compared to the usual durations of performed notes. A COSFIRE filter which is configured with such a prototype, will return multiple strong responses along a note. This fact increases the chances of getting a strong response to the desirable position which in this case is a performed note.

We denote by $P = \{(f_i, \rho_i) \mid i = 1, \ldots, n\}$ a COSFIRE filter that is selective for a given prototype of n points. Each point i of the prototype is described by a pair (f_i, ρ_i), where f_i is the frequency of the note at position (temporal shift) ρ_i with respect to the midpoint of the prototype, where $\rho_i = i - (n+1)/2$. For instance, the COSFIRE filter that is configured by the note "A" of (n=) 5 points, results in the following set P:

$$P = \left\{ \begin{array}{l} (f_1 = 220,\ \rho_1 = \text{-2}), \\ (f_2 = 220,\ \rho_2 = \text{-1}), \\ (f_3 = 220,\ \rho_3 = 0), \\ (f_4 = 220,\ \rho_4 = 1), \\ (f_5 = 220,\ \rho_5 = 2) \end{array} \right\}$$

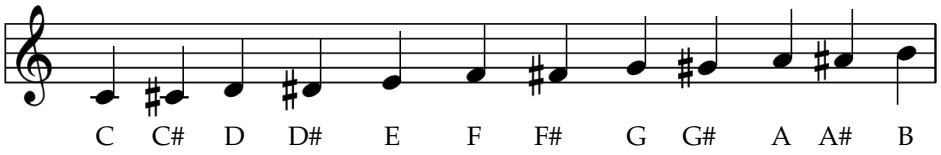

Figure 6.5: The twelve notes of the Western music theory. The Western classical notation is shown first below the notes.

We configure 12 COSFIRE filters that are selective for the third octave of the 12 notes of the Western music theory as illustrated in Fig. 6.5. The frequencies in Hz of those 12 notes are the following: 130.8, 138.6, 146.8, 155.6, 164.8, 174.6, 185, 196, 207.7, 220, 233.1, 246.9, 261.6, 277.2, 293.7, 311.1, 329.6, 349.2.

6.3.3 Response of a COSFIRE filter

We denote by $r_P(t)$ the response of a COSFIRE filter to a signal S at time t. We compute it by taking the geometric mean of the similarity values, which we obtain by a Gaussian kernel function, between the preferred fundamental frequencies f_i and the corresponding frequencies in the concerned neighbourhood.

$$r_P(t) = \left(\prod_{i=1}^{|P|} \exp \frac{-(f_i - S_{t+\rho_i})^2}{2\sigma^2} \right)^{\frac{1}{|P|}}, \sigma = \sigma_0 + \alpha(|\rho_i|) \tag{6.1}$$

where $\sigma_0 = 0.5$ and $\alpha = 0.1$ are constant values that we set experimentally. In this way, the tolerance with respect to the preferred frequency increases with increasing distance from the support center of the COSFIRE filter at hand.

The signal in Fig. 6.6b shows the fundamental frequency of an original note "A" of 92 audio frames covering frequencies in the range between 218 and 222 Hz and we use it as a test signal. The COSFIRE filter P is applied in every position of the test signal and the response is shown in 6.6b. The maximum value of the response is achieved only at the point where the original note has the same values as the prototype. High responses are also achieved for signals that are similar to the prototype.

6.3.4 Post processing

In Fig. 6.7a we show a part of a pitch track that was extracted from one of the songs in our data set. Below it, we illustrate the responses of three COSFIRE filters that are selective for the notes B, C and C# and are shown with thin solid lines.

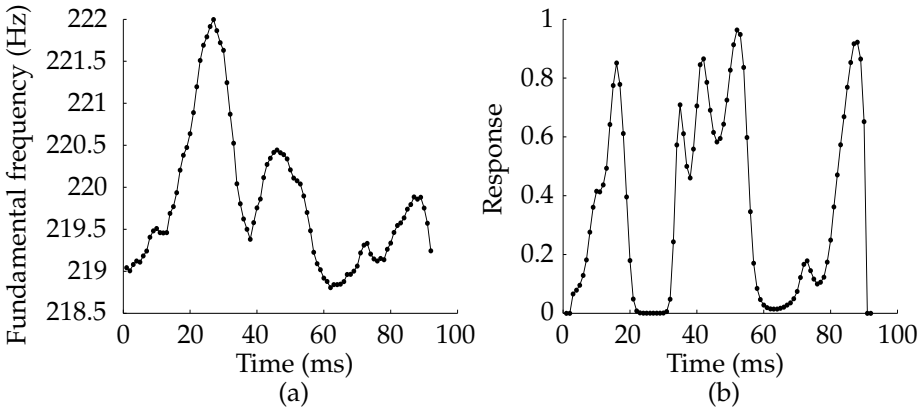

Figure 6.6: (a) Fundamental frequency of an original ornamented note that is used as a test signal. (b) The response of the A-selective COSFIRE filter when applied to the test signal.

Then, we binarize the responses using an absolute threshold of 0.01. We call a unit the consecutive binarized responses that have responses of 1. Such units are shown Fig. 6.7 (b-d) with dashed lines. In case during a time interval between two units there are no responses from other COSFIRE filters, we set all responses between those two units to 1. There is only one such an example in Fig. 6.7 which is obtained by the C-selective filter. If there is temporal overlap between units of different filters we only keep the unit with the longest duration.

Then we transform the binarized responses signals for an input audio signal into a symbolic representation of the sequence that every unit appears. In the example presented in Fig. 6.7 the symbolic sequence is [C.217, B.16, C.6, C#.13, C.6, B.11, C.5 C#.4]. The letters represent the names of the notes of the Western music and the number following the dot is the duration of each unit in number of frames. A vibrato is present at the second half with an alteration of short semitones around the note "C" and it is emphasized in Fig.6.7a.

6.3.5 Ornamentation detection

In order to detect any type of ornamentation we first set a time threshold on the duration of every unit of the symbolic representation. The units that exceed the threshold are considered as notes and the remaining units are marked as parts of ornamentations. Typically, there is a sequence of such short units that consists of the entire ornamentation. In the example shown in Fig. 6.7 we present a note whose second half is characterized by a vibrato. The evolution of the vibrato is emphasised in the pitch track. The first unit which belongs to the note "C" has a duration of 217 time frames while the remaining 7 units have duration of 16, 6, 13, 6, 11, 5 and 4

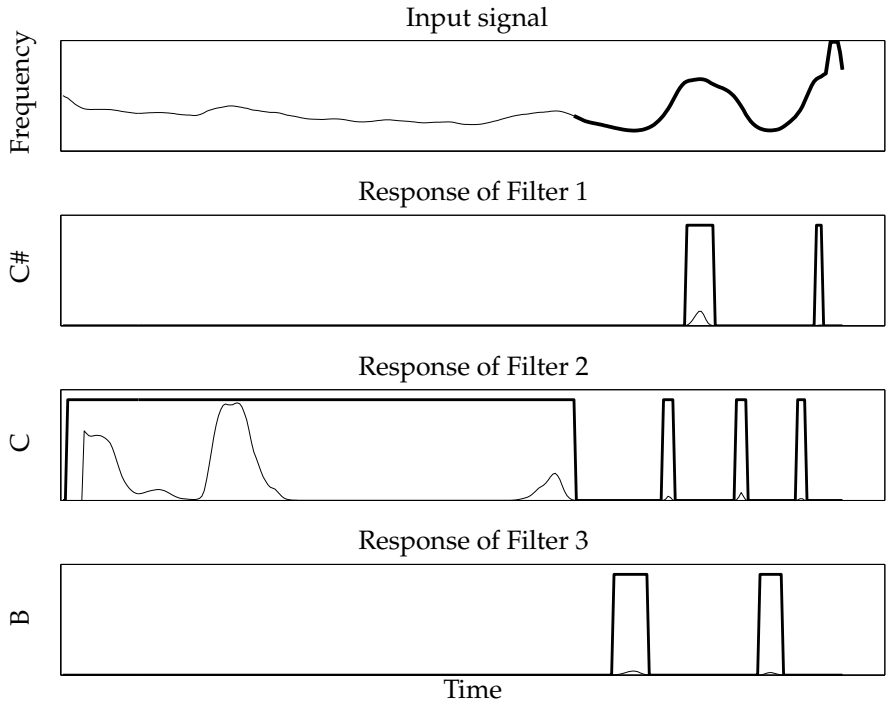

Figure 6.7: Post processing procedure. (a) A part of a pitch track. The automatic detection and recognition of an ornamentation is emphasized in the second half of the signal. (b-d) The responses of three filters that were configured to be selective for the three Western notes C#, C and B. The binarized responses are shown with dashed lines. The thick line in (c) shows how consecutive units of a filter response are connected.

time frames. If we set a threshold of 30 time frames, the first unit will be considered as a note and the following sequence of 7 short notes will be considered to form an ornamentation.

6.3.6 Classification: single-note and multi-note ornamentations.

Each of the notes in Western music can be represented with numbers. For instance, the note A is always the first, hence it is represented with the number 1. Therefore, the symbolic representation shown in the example in Fig. 6.7 can be transformed as [4.217, 3.16, 4.6, 5.13, 4.6, 3.11, 4.5 5.4] where the number preceding the dot repre-

sents the respective note. In this example, the sequence of the notes that forms an ornamentation is: [3, 4, 5, 4, 3, 4, 5]. The classification of an ornamentation into a single-note or multi note is done with counting its peaks and its valleys. If an ornamentation has more than one peak or more than one valley then it is classified as multi-note. The ornamentation shown in Fig. 6.7 has two peaks and one valley and therefore it is classified as a multi-note ornamentation.

If an ornamentation is single-note, then we use four additional rules to classify it into four sub-categories shown in Fig. 6.2. The linear positive and the linear negative do not have peaks or valleys in the sequence and we use the sign of the derivative to decide. For instance, a linear positive ornamentation is detected by the symbolic sequence: [3, 4, 5]. The glissando positive has only one peak while the glissando negative has only one valley. The two subcategories of the multi-note are the vibrato positive and the vibrato negative. They are identified if the ornamentation starts with a positive direction or a negative direction respectively. For example, a vibrato positive is described by the symbolic sequence: [3, 4, 3].

6.4 Experiments and results

6.4.1 Data set

We created a data set of five Cypriot folk songs with total duration of 403 seconds containing 428 ornamentations. They are monophonic singing voice recordings encoded with 44.1kHz sampling frequency. We refer as the ground truth data the positions that are around in the middle of an ornamentation. They are manually annotated by the author of this thesis who is an experienced musician. The list of the songs used to validate our method is given in Table 6.1. From the 428 ornamentations, 270 are single-note and 158 are multi-note. The data set is available online[1].

6.4.2 Results

The results are summarised in Table 6.2 in terms of precision, recall and F-measure. The manual annotation was done by setting a marker in the middle of every ornamentation. Then, we consider a true positive when this marker lies anywhere between the predicted start and end positions of an ornamentation. The false positives are considered when there is no pre-annotated marker between the predicted start and end positions and we count the false negatives when there is a marker with no predictions around it.

In Fig. 6.8 we illustrate a precision-recall plot by changing the time threshold that is used to identify ornamentations from notes. The value of the precision and

[1] https://www.cs.ucy.ac.cy/projects/folk/

Table 6.1: The list of songs used for the validation of our method. Information about the duration, the number of ornamentations and their type is included.

	Manes	Agapisa	Panw xorio	Giallourika	Sousa	Sum
Duration (ms)	84.6	112.3	78.7	74.9	53.3	403.8
Ornamentations (#)	84	123	81	85	55	428
Single-note	56	85	46	50	33	270
Linear positive	9	24	20	10	6	50
Linear negative	7	14	6	17	6	50
Glissando positive	32	42	15	20	16	131
Glissando negative	2	5	5	3	2	17
Multi note	28	38	35	35	22	158
Vibrato positive	23	29	20	32	12	116
Vibrato neg	5	8	15	3	10	41

Table 6.2: Results. The Ornamentation in row 1 contain results that are made of the results in subsequent lines.

	P	R	F-Score
Ornamentations	0.83	0.71	0.76
Single	0.7	0.56	0.63
Single 1	0.7	0.4	0.52
Single 2	0.63	0.62	0.62
Single 3	0.71	0.45	0.55
Single 4	0.29	0.65	0.4
Multi	0.61	0.35	0.45
Multi 1	0.61	0.48	0.54
Multi 2	0.32	0.76	0.45

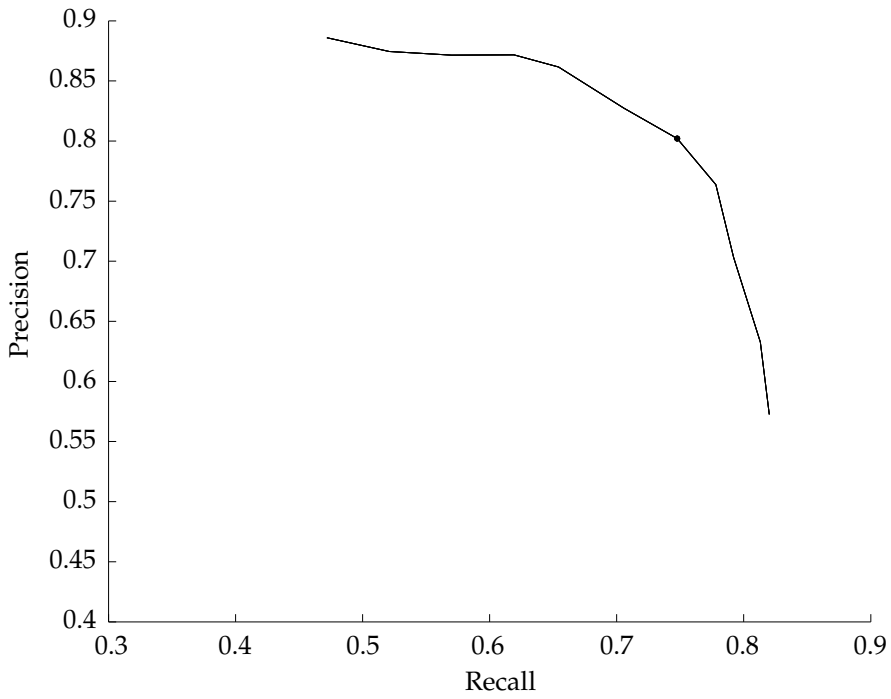

Figure 6.8: Precision-recall plot obtained by varying the time threshold used to distinguish notes from parts of ornamentations.

recall when F-measure reaches maximum is indicated with a dot marker and it has a value of 0.76. We obtain the same results when we used COSFRE filters selective for note signals of length 5, 11, 15, 19 and 21ms.

6.5 Discussion and conclusions

The COSFIRE filters are sensitive to amplitude tolerance, which is essential for non-stationary signals such as the singing voice. They are conceptually simple and easy to implement. They have been successful in pattern recognition of images in several applications including traffic sign detection and recognition (Azzopardi and Petkov 2013). They can be used for other applications such as note segmentation since they are able to capture note changes. The modelling of the pitch track with COSFIRE filters can also be used for the identification of repeating melodies.

In order to compare the proposed 1D COSFIRE filters with an already established method for the identification of similar signals, we cross correlate the same

12 prototypes with the pitch tracks of our data set. We observe that this system returns similar values for all the 12 prototypes and therefore it is not effective for ornamentation detection. Also, the results shown in Table 6.2 are comparable to the ones reported in (Casey and Crawford 2004). Even though the database used in (Casey and Crawford 2004) is different than the one used in this study, we consider that our method is less complex since we avoid the note segmentation which is a challenging task itself. This study is a preliminary work on the ornamentation detection and classification. The classification of the multi ornamentations did not yield significant results, although this is due to the complexity of the problem. We aim to improve the classification stage that is currently rule-based to more sophisticated machine learning techniques. In future work we will attempt to compare our method with other databases used in previous work of other people. The results will be reported in another study. The above, demonstrate that there is a lot of work to be done in this research topic.

The contribution of our work is three-fold. First, it avoids note segmentation which is still a challenging and complex problem. Second, it uses only one feature, the fundamental frequency and third it is effective with complex musical signals such as singing voice. The use of the pitch track as the main input feature to our method contributes to a system that is fast and less complex as compared to other methods (Casey and Crawford 2004, Gainza and Coyle 2007, Köküer et al. 2014). The results that we obtain for the ornamentation detection and recognition are promising and a validation of additional folk music will be done in another study.

Chapter 7

Summary and Outlook

7.1 Summary

This thesis has been organised in two parts. The major objective of the first part, is to use machine learning techniques for the task of the identification of the chromosomal abnormalities during the first trimester of pregnancy. More specifically, classifiers such as Artificial Neural Networks (ANNs), support vector machines and k-nearest neighbours had been applied to a large dataset of pregnant women that underwent for a pre-natal screening. The results of this study are presented in Chapter 2. Currently, in the literature it has been reported that a statistical mixture model (SMM) is used as a classifier to estimate a risk for T21. In this thesis it is shown that ANNs achieve better results than the SMM in both the diagnostic rate (DR) of the T21 at a lower false positive rate (FPR). In addition to this, in the proposed methodology, other chromosomal abnormalities (OCA) such as trisomies 13 and 18, triploidy and Turner syndrome are identified. In the literature, the identification of the OCA is not clearly reported and the commercial use of the pre-natal diagnosis is done only for T21. In the same chapter, the data are visualized and analyzed with standard statistical methods.

In addition to the goal of identifying the fetal chromosomal abnormalities, other technical questions were addressed in Chapter 3. In the literature, it is shown that biochemical parameters from the blood test improve their seperability when they are normalized with the multiples of their medians. Several experiments were held in order to explore the possible use of raw data for training the ANNs. Furthermore, in the same Chapter 3, clustering techniques were used to split the normal (euploid) cases in a number of subclusters and representative instances were collected around the prototypes of the k-means for creating a balanced training set. We showed that better results are achieved for the DR of the OCA, but the difference of the results for the DR of T21 for the balanced and the imbalanced training sets is not significant.

In Chapter 4, a two stage approach for the identification of fetal chromosomal abnormalities was proposed. Following the findings from Chapter 3, it is shown that all pregnant women can perform at a first stage a pre-natal diagnosis test using 4 parameters from the ultrascan and the blood test. From the results in stage 1, all the T21 (100% TPR) and 77% of the OCA are identified with a cost of a relatively

high FPR of about 20%. In stage 2, all the cases that are ranked positive in stage 1 are advised to proceed with an additional examination to get the values of two additional parameters, the ductus venosus and the tricuspid flow. The output of stage 2 is separated in "no risk", "moderate risk" and "high risk" areas. The cases that are ranked as "moderate risk" are suggested to perform another non-invasive test, called cell-free fetal DNA, while the cases in "high risk" are suggested to proceed with one of the invasive tests, the amniocentesis or the chorionic villus sampling. In conclusion, in this work, we show that ANNs perform significantly better than other existing non-invasive tests.

In the second part of this thesis, two applications in the field of computational ethnomusicology had been presented. The first application identifies important motifs in 1D signals from melodic sequences and converts a song into a symbolic representation that is based on the similarity and the repetition of the appearing motifs. We introduce the concept of the two-layers COSFIRE approach that are configured using properties from the fundamental frequency of the audio signal and are tuned with several parameters. The COSFIRE filters have been effective in 2D signals and particularly in images (Azzopardi and Petkov 2013, Azzopardi and Petkov 2014). We adapted the 2D COSFIRE filters for 1D signals and we report their effectiveness in a benchmark dataset of 38 songs (X motifs). The results are measured in terms of precision and recall for the motif identification. For the symbolic representation we use the minimum hamming distance between the string given in the ground truth and the resulting string from the application. We compare our method with other existing methods such as dynamic time warping (DTW) and cross correlation. The COSFIRE approach as shown in Chapter 5 are more effective and more efficient than the other existing methods that are compared with the same dataset.

In Chapter 6, the task of the identification of ornamentations in folk singing songs is addressed. We have used the theoretical frequencies of the third octave of the Western music theory to configure 12 COSFIRE filters that are applied to the fundamental frequency of the audio signal of five Cypriot folk songs with total duration of 403 seconds containing 428 ornamentations. A cut-off value is used to binarize the response signals of the 12 COSFIRE filters and dynamic programming is applied for the identification of ornamentations. The proposed method classifies a detected ornamentation in one of six categories that are based on the literature and in music theory. We applied the same procedure by using another similarity distance, the cross correlation, and is found that COSFIRE filters perform significantly better.

7.2 Outlook

The work presented in this thesis can be extended in several directions. It is discussed in the scientific community through the fetal medicine foundation world congress and other relevant conferences that the use of characteristics coming from the father could possibly contribute to the detection of chromosomal abnormalities. So far, none of the biological or other features of the father is used in the existing methods for the pre-natal diagnosis test for chromosomal abnormalities. Essentially, this is an open question that has practical obstacles to be researched, such as the doubt of the identity of the father.

The pre-eclampsia is a serious and at some cases fatal disease that can be developed to the mother during pregnancy. It is reported in the literature that the symptoms of pre-eclampsia are related to a fetus with chromosomal abnormality. For instance, the two biomarkers that are extracted from the maternal blood respond in a similar manner in both situations either the evolution of pre-eclampsia, either a fetal chromosomal abnormality, either both. After several meetings with the collaborating doctors, we have made a model for future work that will use the parameters of the pre-natal examinations and will predict whether a case is developing pre-eclampsia, or carries a fetus with chromosomal abnormality.

Considering the second part of this thesis, it is shown that the use of COSFIRE filters in 1D ethnomusicological sequences is promising for solving several tasks such as motif repetition and ornamentation detection. However, more work needs to be done in order to improve the initial segmentation that is currently done manually. Moreover, in both methods that are proposed in Chapters 5 and 6, the parameters for tuning and optimizing the COSFIRE filters have to be done using a greed search approach. Automated setup of the parameters can be achieved, depending on the task. For instance, the parameters for the amplitude tolerance can be set according to the mean and the standard deviation of the input signal.

From the experiments done in Chapter 6, it is shown that the filter approach is promising for note-level segmentation. However, some limitations of this work are identified. First, the dataset used is narrowed into monophonic songs by male performers. In order to capture the melodic characteristics of these songs, we configured COSFIRE filters that are selective only for the frequency range of a male singer. It would be interesting to widen the dataset by including female singers and make the method to be selective in the entire frequency range of the Western music theory. Finally, the systems in Part II can be improved by extracting the fundamental frequency of polyphonic music. This can be done using existing algorithms such as MELODIA (Salamon and Gómez 2012).

Another application where COSFIRE filters could possibly become applicable, would be a pairwise similarity measure between two entire songs. For this task, a lot of work has been done in the MIR community and it has a particular interest

in the field of computational ethnomusicology. From the properties of the COSFIRE filters as has been designed to allow temporal and frequency tolerance, it is expected that they can achieve good results but not necessarily better than other methods.

Chapter 8

Discussion

The identification of the fetal chromosomal abnormalities in the early stage of pregnancy is a difficult task. One needs to deal with a big medical dataset where the ratio between the populations of the normal and the abnormal cases is highly imbalanced. Moreover, in the data collection procedure, some of the markers are manually annotated by the doctor. One example of such marker is the fetal crown rump length, a distance measurement between the upper and the lower body of the fetus. It can frequently become complicated to make an accurate measurement looking at the ultrasonographic screening. In some occasions the doctor has to enforce the fetus into another position and look up again. From the statistical analysis that is presented in Chapter 2 and from the figures in the Appendix, we observed a higher variance compared to other markrers such as the PAPP-A and the β-hCG. Therefore, even though the actual information of the statistical properties of the data is not used to optimize the ANNs that are used for classification, the statistical analysis that is presented in Part I was done in order to understand, confirm and visualize any statistical errors or other observations.

Nevertheless, from a practical point of view, some limitations of the work done in Part I are identified. For instance, the most common questions that a researcher has to give to himself, are which classifier should I use, and how to optimize its parameters? It is indeed a hard question to answer and this is where the work presented in this thesis lacks. In this work, I have followed a heuristic approach to test the most commonly used classifiers for similar medical applications and I chose to continue my work with the one that yielded the best results, being the ANNs as shown in Chapter 2. The optimization of the parameters of each classifier is done in a similar manner. Therefore, the statement that "the forward neural systems are proved to be the most suitable from the point of view of satisfactory generalization and diagnostic yield for such predictive systems" may not include a clear justification on how the specific architecture was selected as the best one.

As stated above, the number of the normal cases in our dataset is significantly higher than the number of the abnormal cases. In Chapter 3, I propose a method for creating a subset of the normal cases for training, that are chosen around some k-means prototypes and can be think of representative cases of the entire population. Here, a similar question with the one stated above is which one of the methods that are proposed in the literature are the most suitable for dealing with the class imbalanced problem of this particular dataset? Here is another point that it possibly needs

better clarification for the use of the k-means algorithm, as described in Chapter 3. It is based on the assumption that the normal cases are consisted with a number of subclasses and one way to separate them is to use an unsupervised method, that could be any other method. Nevertheless, in this thesis contributes to the fact that there is an indication that data reduction of the normal cases in the training set is preferred or needed.

What differentiates Part I with Part II is essentially the nature of the data. While in Part I I deal with a classification problem where most or all of the variables are independent, in Part II in the musicological data there is the factor of time playing a significant role to the methodology which has to be approached. For the identification of several musical higher level features, I introduced an adaptation of the COSFIRE filters that are well proven to be effective for the extraction of spatial patterns. The adaptation of a filter that is implemented for 2D and 3D to 1D signal can be considered simple. However, time signals typically include particularities in terms of frequency, signal/noise ratio, scale of the pattern, and overlap between several pattern instances. The results in Part II show the wider generalisability of COSFIRE filters beyond spatial data. In the context of the adaptation of the COSFIRE filters in a lower dimensional space, one may question what would be the performance of such filters for tasks with higher dimensionality. What distinguishes COSFIRE filters with other methods such as Gaussian Mixture Models, is that they incorporate tolerance between the expected values and the position with respect to the filter support. So far, they are used in applications where data points have some type of order. In image processing, the polar coordinates and the distance between the filter support and the point of interest are important. In the ethnomusicological data and the applications that are presented in this thesis, the time has similar importance. The standard deviation of the gaussian function between the expected values and the test signal is increased, allowing tolerance in the amplitude. In a second stage, a blurring function is used for tolerating the similarity measure in temporal space. Therefore, COSFIRE filters could possibly work for applications such as videos or motion tracking that require more than 3 dimensional data.

One question here is whether would ANNs be effective for the identification of repeating patterns and ornamentation detection as described in Part II and similarly, if COSFIRE filters would successfully respond to chromosomal abnormalities, given the feature vector of an unknown case. Regarding the first question, ANNs and generally machine learning techniques are applied and reported in the literature for many applications. For instance, segmentation and annotation tasks for musical information such as musical or vocal parts, note onset, genre detection and others can be seen as classification problems and therefore can be approached by machine learning methods. Applications for timbre analysis, instrument identification, cover and genre detection, typically a number of global features are extracted and each song is characterized by a number of descriptors. For the second ques-

tion, whether COSFIRE filters could be applied for the detection of chromosomal abnormalities, we need to state some hypotheses. As mentioned above, COSFIRE filters are applicable in applications where the data points that are away from the filter support have some type of order. In the case of the chromosomal abnormalities dataset, the markers are independent from the concept of time or position and therefore COSFIRE filters would possibly become similar to other methods. The most fundamental difference between ANNs and COSFIRE filters is the fact that a COSFIRE filter is a feature detector and not a classification model. In the applications reported in Chapters 5 and 6, the COSFIRE filters can be considered as a high nonlinear similarity function between a test pattern and the prototype (preferred) pattern.

Bibliography

Abe, H. and Yamaguchi, T.: 2005, Implementing an integrated time-series data mining environment-a case study of medical kdd on chronic hepatitis, *1st international conference on complex medical engineering (CME2005)*.

Al-Shayea, Q. K.: 2011, Artificial neural networks in medical diagnosis, *International Journal of Computer Science Issues* **8**(2), 150–154.

Ashoor, G., Syngelaki, A., Wagner, M., Birdir, C. and Nicolaides, K. H.: 2012, Chromosome-selective sequencing of maternal plasma cell–free dna for first-trimester detection of trisomy 21 and trisomy 18, *American journal of obstetrics and gynecology* **206**(4), 322–e1.

Aucouturier, J.-J. and Sandler, M.: 2002, Finding repeating patterns in acoustic musical signals : Applications for audio thumbnailing, *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Espoo, Finland.

Azzopardi, G. and Petkov, N.: 2013, Trainable COSFIRE filters for keypoint detection and pattern recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 490–503.

Azzopardi, G. and Petkov, N.: 2014, Ventral-stream-like shape representation: from pixel intensity values to trainable object-selective COSFIRE models, *Frontiers in Computational Neuroscience* **8**.

Bartsch, M. A. and Wakefield, G. H.: 2005, Audio thumbnailing of popular music using chroma-based representations, *IEEE Transactions on Multimedia* **7**, 96–104.

Benetos, E., Kotti, M. and Kotropoulos, C.: 2006, Applying supervised classifiers based on non-negative matrix factorization to musical instrument classification, *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME*, Toronto, Ontario, Canada, pp. 2105–2108.

Bianchi, D. W., Flint, A. F., Pizzimenti, M. F., Knoll, J. and Latt, S. A.: 1990, Isolation of fetal dna from nucleated erythrocytes in maternal blood, *Proceedings of the National Academy of Sciences* **87**(9), 3279–3283.

Biehl, M., Hammer, B. and Schneider, P.: 2006b, Matrix learning in learning vector quantization, *Insitute of Informatics, Clausthal University of Technology* pp. 06–14.

Boenn, G.: 2007, Automated quantisation and transcription of ornaments from audio recordings.

Boughorbel, S., Al-Ali, R. and Elkum, N.: 2016, Model comparison for breast cancer prognosis based on clinical data, *PloS one* **11**(1), e0146413.

Bozkurt, B.: 2008, An automatic pitch analysis method for turkish maqam music, *Journal of New Music Research* **37**, 1–13.

Brambati, B., Simoni, G., Danesino, C., Oldrini, A., Ferrazzi, E., Romitti, L., Terzoli, G., Rossella, F., Ferrari, M. and Fraccaro, M.: 1985, First trimester fetal diagnosis of genetic disorders: clinical evaluation of 250 cases., *Journal of medical genetics* **22**(2), 92–99.

Butler, D.: 1989, Describing the perception of tonality in music: A critique of the tonal hierarchy theory and a proposal for a theory of intervallic rivalry, *Music Perception: An Interdisciplinary Journal* **6**, 219–241.

C, L. and Chang, J.: 2012, Enhancing support vector machine-based speech/music classification using conditional maximum a posteriori criterion, *IET Signal Processing* **6**(4), 335–340.

Camilleri, T. A., Camilleri, K. P. and Fabri, S. G.: 2014, Automatic detection of spindles and k-complexes in sleep eeg using switching multiple models, *Biomedical Signal Processing and Control* **10**, 117–127.

Casey, M. and Crawford, T.: 2004, Automatic location and measurement of ornaments in audio recordings, *Proceedings of the 5th International Symposium of Music Information Retrieval (ISMIR)*.

Chiu, R. W., Akolekar, R., Zheng, Y. W., Leung, T. Y., Sun, H., Chan, K. A., Lun, F. M., Go, A. T., Lau, E. T., To, W. W. et al.: 2011, Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma dna sequencing: large scale validity study, *Bmj* **342**, c7401.

Crochemore, M., Iliopoulos, C. S., Lecroq, T. and Pinzon, Y. J.: 2001, Approximate string matching in musical sequences, *Prague Stringology Conference'01*, Prague, Czech Republic, pp. 26–36.

Dannenberg, R. B., Birmingham, W. P., Tzanetakis, G., Meek, C., Hu, N. and Pardo, B.: 2004, The musart testbed for query-by-humming evaluation, *Computer Music Journal* **28**(2), 34–48.

Dannenberg, R. B. and Hu, N.: 2004, Understanding search performance in query-by-humming systems., *ISMIR*.

de Cheveigné, A. and Kawahara, H.: 2002, Yin, a fundamental frequency estimator for speech and music, *J Acoust Soc Am* **111**, 1917–1930.

Dinov, I. D., Heavner, B., Tang, M., Glusman, G., Chard, K., Darcy, M., Madduri, R., Pa, J., Spino, C., Kesselman, C. et al.: 2016, Predictive big data analytics: A study of parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations, *PloS one* **11**(8), e0157077.

Dorken, E. and Nawab, S. H.: 1994, Improved musical pitch tracking using principal decomposition analysis, *In International Conference on Acoustics, Speech and Signal Processing* **2**, 217–220.

Duchene, F., Garbay, C. and Rialle, V.: 2005, Apprentissage de motifs temporels, multidimensionnels et heterogenes: Application a la telesurveillance medicale, *Proceedings of conference francophone sur lÁpprentissage Automatique*, Tregastel, France.

Dugoff, L. et al.: 2010, First-and second-trimester maternal serum markers for aneuploidy and adverse obstetric outcomes, *Obstetrics & Gynecology* **115**(5), 1052–1061.

Dutta, S. and Murthy, H. A.: 2014, Discovering typical motifs of a raga from one-liners of songs in carnatic music., *ISMIR*, pp. 397–402.

Ehrich, M., Deciu, C., Zwiefelhofer, T., Tynan, J. A., Cagasan, L., Tim, R., Lu, V., McCullough, R., McCarthy, E., Nygren, A. O. et al.: 2011, Noninvasive detection of fetal trisomy 21 by sequencing of dna in maternal blood: a study in a clinical setting, *American journal of obstetrics and gynecology* **204**(3), 205–e1.

Enzensberger, C., Pulvermacher, C., Degenhardt, J., Kawacki, A., Germer, U., Gembruch, U., Krapp, M., Weichert, J. and Axt-Fliedner, R.: 2012, Fetal loss rate and associated risk factors after amniocentesis, chorionic villus sampling and fetal blood sampling., *Ultraschall in der Medizin (Stuttgart, Germany: 1980)* **33**(7), E75–9.

Er, O., Cetin, O., Bascil, M. S. and Temurtas, F.: 2016, A comparative study on parkinson's disease diagnosis using neural networks and artificial immune system, *Journal of Medical Imaging and Health Informatics* **6**(1), 264–268.

Er, O., Yumusak, N. and Temurtas, F.: 2010, Chest diseases diagnosis using artificial neural networks, *Expert Systems with Applications* **37**(12), 7648–7655.

Eronen, A.: 2007, Chorus detection with combined use of mfcc and chroma features and image processing filters, *In Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07*, Bordeaux, France.

Fang, R., Pouyanfar, S., Yang, Y., Chen, S.-C. and Iyengar, S.: 2016, Computational health informatics in the big data age: A survey, *ACM Computing Surveys (CSUR)* **49**(1), 12.

Ferreira, P., Azevedo, P., Silva, C. and Brito, R.: 2006, Mining approximate motifs in time series, *Proceedings of the 9th International Conference on Discovery Science*, Barcelona, Spain, pp. 89–101.

Gainza, M. and Coyle, E.: 2007, Automating ornamentation transcription, *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Vol. 1, IEEE, pp. I–69.

Giannoulis, D. and Klapuri, A.: 2013, Musical instrument recognition in polyphonic audio using missing feature approach., *IEEE Transactions on Audio, Speech & Language Processing* **21**, 1805–1817.

Gómez, E.: 2006, *Tonal Description of Music Audio Signals*, PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain.

Goto, M.: 2006, A chorus section detection method for musical audio signals and its application to a music listening station, *Trans. Audio, Speech and Lang. Proc.* **14**, 1783–1794.

Hayashi, Y. and Setiono, R.: 2002, Combining neural network predictions for medical diagnosis, *Computers in biology and medicine* **32**(4), 237–246.

Hillewaere, R., Manderick, B. and Conklin, D.: 2009, Global feature versus event models for folk song classification, *10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan.

Hoens, T. R., Chawla, N. V., He, H. and Ma, Y.: 2013, Imbalanced datasets: from sampling to classifiers, *Imbalanced Learning: Foundations, Algorithms and Applications. Wiley* pp. 43–59.

Howell, P., West, R. and Cross, I.: 1991, *Representing Musical Structure*, number 5 in *Cognitive Science Series*, Academic.

Hu, N., Dannenberg, R. B. and Tzanetakis, G.: 2003, Polyphonic audio matching and alignment for music retrieval, *Computer Science Department* p. 521.

Huggon, I., DeFigueiredo, D. and Allan, L.: 2003, Tricuspid regurgitation in the diagnosis of chromosomal anomalies in the fetus at 11–14 weeks of gestation, *Heart* **89**(9), 1071–1073.

Huq, A., Cartwright, M. and Pardo, B.: 2010, Crowdsourcing a real-world on-line query by humming system, *Proceedings of the Sixth Sound and Music Computing Conference (SMC 2010)*.

Ishizuka, K. and Nakatani, T.: 2006, A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition, *Speech Communication* **48**(11), 1447–1457.

Ishwar, V., Dutta, S., Bellur, A. and Murthy, H. A.: 2013, Motif spotting in an alapana in carnatic music., *ISMIR*, pp. 499–504.

Janssen, B., van Kranenburg, P. and Volk, A.: 2015, A comparison of symbolic similarity measures for finding occurrences of melodic segments, *Proceedings of the 16th ISMIR Conference, Málaga, Spain, October 26-30, 2015*, ISMIR press, pp. 659–665.

Jehan, T.: 2005, *Creating Music by Listening*, PhD thesis, Massachusetts Institute of Technology, Massachusetts, USA.

Jeon, W., Ma, C. and Cheng, Y. M.: 2009, An efficient signal-matching approach to melody indexing and search using continuous pitch contours and wavelets., *ISMIR*, pp. 681–686.

Kagan, K. O., Wright, D. and Nicolaides, K.: 2015, First-trimester contingent screening for trisomies 21, 18 and 13 by fetal nuchal translucency and ductus venosus flow and maternal blood cell-free dna testing, *Ultrasound in Obstetrics & Gynecology* **45**(1), 42–47.

Kagan, K., Staboulidou, I., Syngelaki, A., Cruz, J. and Nicolaides, K.: 2010, The 11–13-week scan: diagnosis and outcome of holoprosencephaly, exomphalos and megacystis, *Ultrasound in Obstetrics & Gynecology* **36**(1), 10–14.

Kagan, K., Valencia, C., Livanos, P., Wright, D. and Nicolaides, K.: 2009, Tricuspid regurgitation in screening for trisomies 21, 18 and 13 and turner syndrome at 11+ 0 to 13+ 6 weeks of gestation, *Ultrasound in Obstetrics & Gynecology* **33**(1), 18–22.

Kagan, K., Wright, D., Baker, A., Sahota, D. and Nicolaides, K.: 2008, Screening for trisomy 21 by maternal age, fetal nuchal translucency thickness, free beta-human chorionic gonadotropin and pregnancy-associated plasma protein-a, *Ultrasound in Obstetrics & Gynecology* **31**(6), 618–624.

Kagan, K., Wright, D., Spencer, K., Molina, F. and Nicolaides, K.: 2008, First-trimester screening for trisomy 21 by free beta-human chorionic gonadotropin and pregnancy-associated plasma protein-a: impact of maternal and pregnancy characteristics, *Ultrasound in Obstetrics & Gynecology* **31**(5), 493–502.

Keogh, E., Lin, J., Lee, S.-H. and Van Herle, H.: 2006, Finding the most unusual time series subsequence: Algorithms and applications, *Knowl. Inf. Syst.* **11**, 1–27.

Khan, I. Y., Zope, P. and Suralkar, S.: 2013, Importance of artificial neural network in medical diagnosis disease like acute nephritis disease and heart disease, *International Journal of Engineering Science and Innovative Technology (IJESIT)* **2**(2), 210–217.

Klapuri, A. P.: 2004, Automatic music transcription as we know it today, *Journal of New Music Research* **33**, 269–282.

Klement, W., Wilk, S., Michalowski, W., Farion, K. J., Osmond, M. H. and Verter, V.: 2012, Predicting the need for ct imaging in children with minor head injury using an ensemble of naive bayes classifiers, *Artificial intelligence in medicine* **54**(3), 163–170.

Köküer, M., Jancovic, P., Ali-MacLachlan, I. and Athwal, C.: 2014, Automated detection of single-and multi-note ornaments in irish traditional flute playing., *ISMIR*, pp. 15–20.

Kolomogoroff, A.: 2013, *Grundbegriffe der wahrscheinlichkeitsrechnung*, Vol. 2, Springer-Verlag.

Kotsifakos, A., Papapetrou, P., Hollmén, J., Gunopulos, D., Athitsos, V. and Kollios, G.: 2012, Hum-a-song: a subsequence matching with gaps-range-tolerances query-by-humming system, *Proceedings of the VLDB Endowment* **5**(12), 1930–1933.

Kubat, M., Matwin, S. et al.: 1997, Addressing the curse of imbalanced training sets: one-sided selection, *ICML*, Vol. 97, Nashville, USA, pp. 179–186.

Lee, S., Wu, Z., Dong, M., Tian, X. and Li, H.: 2014, A comparative study of spectral transformation techniques for singing voice synthesis, *Fifteenth Annual Conference of the International Speech Communication Association*.

Levine, D. S.: 2000, *Introduction to neural and cognitive modeling*, Psychology Press.

Lin, J., Keogh, E., Wei, L. and Lonardi, S.: 2007, Experiencing sax: A novel symbolic representation of time series, *Data Min. Knowl. Discov.* **15**, 107–144.

Liu, Q., Cui, X., Chou, Y.-C., Abbod, M. F., Lin, J. and Shieh, J.-S.: 2015, Ensemble artificial neural networks applied to predict the key risk factors of hip bone fracture for elders, *Biomedical Signal Processing and Control* **21**, 146–156.

Maithili, A., Kumari, R. V. and Rajamanickam, M. S.: 2011, Neural networks towards medical diagnosis, *Int. Jour. of Mod. Eng. Res* **1**(1), 57–64.

Maiz, N., Valencia, C., Kagan, K., Wright, D. and Nicolaides, K.: 2009, Ductus venosus doppler in screening for trisomies 21, 18 and 13 and turner syndrome at 11–13 weeks of gestation, *Ultrasound in Obstetrics & Gynecology* **33**(5), 512–517.

Matthews, B. W.: 1975, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, *Biochimica et Biophysica Acta (BBA)-Protein Structure* **405**(2), 442–451.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A. and Tourassi, G. D.: 2008, Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, *Neural networks* **21**(2), 427–436.

McGovern, A., Kruger, A., Rosendahl, D. and Droegemeier, K.: 2006, Open problem: Dynamic relational models for improved hazardous weather prediction, *Proceedings of ICML Workshop on Open Problems in Statistical Relational Learning*, Oregon, USA.

McGovern, A., Rosendahl, D., Kruger, A., Beaton, M., Brown, R. and Droegemeier, K.: 2007, Understanding the formation of tornadoes through data mining, *5th conference on artificial intelligence and its applications to environmental sciences at the American meteorological society*.

Mckinney, M. and Breebaart, J.: 2003, Features for audio and music classification, *Proceedings of the International Symposium on Music Information Retrieval*, pp. 151–158.

Medan, J., Yair, E. and Chazan, D.: 1991, Super resolution pitch determination of speech signals, *IEEE Transactions on Signal Processing* **39**, 40–48.

Muda, L., Begam, M. and Elamvazuthi, I.: 2010, Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques, *CoRR* **abs/1003.4083**.

Müller, M.: 2007, Dynamic time warping, *Information retrieval for music and motion* pp. 69–84.

Müller, M., Grosche, P. and Jiang, N.: 2011, A segment-based fitness measure for capturing repetitive structures of music recordings., *ISMIR*, Citeseer, pp. 615–620.

Myers, C. L., Dunham, M. J., Kung, S.-Y. and Troyanskaya, O. G.: 2004, Accurate detection of aneuploidies in array cgh and gene expression microarray data, *Bioinformatics* **20**(18), 3533–3543.

Neocleous, A. C., Neocleous, C. K., Nicolaides, K. H. and Schizas, C. N.: 2016b, Marker selection for the detection of trisomy 21 using the generalized matrix learning vector quantization, *Submitted for publication, International Joint Conference on Neural Networks*, Anchorage, Alaska.

Neocleous, A. C., Neocleous, C. K., Petkov, N., Nicolaides, K. H. and Schizas, C. N.: 2016, *Prenatal Diagnosis of Aneuploidy Using Artificial Neural Networks in Relation to Health Economics*, Springer International Publishing, pp. 930–934.

Neocleous, A., Panteli, M., Petkov, N. and Schizas, C. N.: 2012, Identification of similarities between the turkish makam scales and the cypriot folk music, *HELINA's 5th National conference*.

Neocleous, A., Petkov, N. and Schizas, C. N.: 2014a, Automated classification in vocal/instrumental parts of folk songs, *7nth Cyprus workshop on signal processing and informatics*.

Neocleous, A., Petkov, N. and Schizas, C. N.: 2014b, Automated segmentation of folk songs using artificial neural networks, *Proceedings of the International Conference on Neural Computation Theory and Applications*, Rome, Italy, pp. 144–151.

Neocleous, C., Nikolaides, K. H., Neokleous, K. C. and Schizas, C. N.: 2010, Artificial neural networks for non-invasive chromosomal abnormality screening of fetuses., *IJCNN*, pp. 1–4.

Neves, J., Guimarães, T., Gomes, S., Vicente, H., Santos, M., Neves, J., Machado, J. and Novais, P.: 2015, Logic programming and artificial neural networks in breast cancer detection, *Advances in Computational Intelligence*, Springer, pp. 211–224.

Nicolaides, K. H.: 2004, Nuchal translucency and other first-trimester sonographic markers of chromosomal abnormalities, *American journal of obstetrics and gynecology* **191**(1), 45–67.

Nicolaides, K. H.: 2005, First-trimester screening for chromosomal abnormalities, *Seminars in perinatology*, Vol. 29, Elsevier, pp. 190–194.

Nicolaides, K. H.: 2011, Screening for fetal aneuploidies at 11 to 13 weeks, *Prenatal diagnosis* **31**(1), 7–15.

Nicolaides, K. H., Azar, G., Byrne, D., Mansur, C. and Marks, K.: 1992, Fetal nuchal translucency: ultrasound screening for chromosomal defects in first trimester of pregnancy., *Bmj* **304**(6831), 867–869.

Nicolaides, K. H., Syngelaki, A., Ashoor, G., Birdir, C. and Touzet, G.: 2012, Noninvasive prenatal testing for fetal trisomies in a routinely screened first-trimester population, *American journal of obstetrics and gynecology* **207**(5), 374–e1.

Nicolaides, K. H., Syngelaki, A., Gil, M. d. M., Quezada, M. S. and Zinevich, Y.: 2013, Prenatal detection of fetal triploidy from cell-free dna testing in maternal blood, *Fetal diagnosis and therapy* **35**(3), 212–217.

Nicolaides, K., Spencer, K., Avgidou, K., Faiola, S. and Falcon, O.: 2005, Multicenter study of first-trimester screening for trisomy 21 in 75 821 pregnancies: results and estimation of the potential impact of individual risk-orientated two-stage first-trimester screening, *Ultrasound in obstetrics & gynecology* **25**(3), 221–226.

Norton, M. E., Brar, H., Weiss, J., Karimi, A., Laurent, L. C., Caughey, A. B., Rodriguez, M. H., Williams, J., Mitchell, M. E., Adair, C. D. et al.: 2012, Non-invasive chromosomal evaluation (nice) study: results of a multicenter prospective cohort study for detection of fetal trisomy 21 and trisomy 18, *American journal of obstetrics and gynecology* **207**(2), 137–e1.

Norton, M. E., Jacobsson, B., Swamy, G. K., Laurent, L. C., Ranzini, A. C., Brar, H., Tomlinson, M. W., Pereira, L., Spitz, J. L., Hollemon, D. et al.: 2015, Cell-free dna analysis for noninvasive examination of trisomy, *New England Journal of Medicine* **372**(17), 1589–1597.

Palomaki, G. E., Kloza, E. M., Lambert-Messerlian, G. M., Haddow, J. E., Neveux, L. M., Ehrich, M., van den Boom, D., Bombard, A. T., Deciu, C., Grody, W. W. et al.: 2011, Dna sequencing of maternal plasma to detect down syndrome: an international clinical validation study, *Genetics in medicine* **13**(11), 913–920.

Panteli, M.: 2011, *Pitch patterns of cypriot folk music between byzantine and ottoman influence*, Master's thesis, Universitat Pompeu Fabra, Barcelona, Spain.

Papageorgiou, E. A., Karagrigoriou, A., Tsaliki, E., Velissariou, V., Carter, N. P. and Patsalis, P. C.: 2011, Fetal-specific dna methylation ratio permits noninvasive prenatal diagnosis of trisomy 21, *Nature medicine* **17**(4), 510–513.

Patel, J. L. and Goyal, R. K.: 2007, Applications of artificial neural networks in medical science, *Current clinical pharmacology* **2**(3), 217–226.

Payne, E.: 2006, Phonetic motifs and the formation of sound structure, *Working Papers in Linguistics* pp. 322–343.

Peeters, G. and Rodet, X.: 2004, A large set of audio feature for sound description (similarity and classification) in the cuidado project, *Technical report*, IRCAM, Paris, France.

Peiszer, E.: 2007, *Automatic audio segmentation: Segment boundary and structure detection in popular music*, Master's thesis, Vienna University of Technology, Vienna, Austria.

Pérez-Ortiz, M., Gutiérrez, P. A., Tino, P. and Hervás-Martínez, C.: 2015, Oversampling the minority class in the feature space.

Qi, Y., Paisley, J. W. and Carin, L.: 2007, Music analysis using hidden markov mixture models, *IEEE Transactions on Signal Processing* **55**, 5209–5224.

Ramalingam, T. and Dhanalakshmi, P.: 2014, Speech/music classification using wavelet based feature extraction techniques, *Journal of Computer Science* **10**(1), 34–44.

Ramirez, R. and Maestre, E.: 2009, A framework for performer identification in audio recordings, *SECOND INTERNATIONAL WORKSHOP ON MACHINE LEARNING AND MUSIC*, p. 6.

Ramirez, R., Maestre, E., Pertusa, A., Gómez, E. and Serra, X.: 2007, Performance-based interpreter identification in saxophone audio recordings, *Circuits and Systems for Video Technology, IEEE Transactions on* **17**(3), 356–364.

Ramirez, R., Perez, A., Kersten, S. and Maestre, E.: 2008, Performer identification in celtic violin recordings., *ISMIR*, pp. 483–488.

Rao, P., Ross, J. C., Ganguli, K. K., Pandit, V., Ishwar, V., Bellur, A. and Murthy, H. A.: 2014, Classification of melodic motifs in raga music with time-series matching, *Journal of New Music Research* **43**(1), 115–131.

Reynolds, T. and Penney, M. D.: 1990, The mathematical basis of multivariate risk screening: with special reference to screening for down's syndrome associated pregnancy, *Annals of Clinical Biochemistry: An international journal of biochemistry in medicine* **27**(5), 452–458.

Robinson, H. and Fleming, J.: 1975, A critical evaluation of sonar crown-rump length measurements, *BJOG: An International Journal of Obstetrics & Gynaecology* **82**(9), 702–710.

Rodrıguez-López, M. E. and Volk, A.: n.d., Selective acquisition techniques for enculturation-based melodic phrase segmentation.

Ross, J. C., Vinutha, T. and Rao, P.: 2012, Detecting melodic motifs from audio for hindustani classical music., *ISMIR*, pp. 193–198.

Ryynänen, M. and Klapuri, A.: 2008, Query by humming of midi and audio using locality sensitive hashing, *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, IEEE, pp. 2249–2252.

Saeidi, R., Sedlak, F., Lee, K. A., Sandberg, J., Hansson-Sandsten, M. and Li, H.: 2012, Low-variance multitaper mfcc features: A case study in robust speaker verification, *Audio, Speech, and Language Processing, IEEE Transactions on* **20**.

Sahoo, T., Cheung, S. W., Ward, P., Darilek, S., Patel, A., Del Gaudio, D., Kang, S. H. L., Lalani, S. R., Li, J., Mcadoo, S. et al.: 2006, Prenatal diagnosis of chromosomal abnormalities using array-based comparative genomic hybridization, *Genetics in Medicine* **8**(11), 719–727.

Salamon, J. and Gómez, E.: 2012, Melody extraction from polyphonic music signals using pitch contour characteristics, *IEEE Transactions on Audio, Speech, and Language Processing* **20**(6), 1759–1770.

Sandler, M. and Aucouturier, J.-J.: 2001, Segmentation of musical signals using hidden markov models, *Audio Engineering Society Convention 110*, Amsterdam, The Netherlands.

Schizas, C. N. and Pattichis, C. S.: 1997, Learning systems in biosignal analysis, *BioSystems* **41**(2), 105–125.

Schmorl, G.: 1893, *Pathologisch-anatomische untersuchungen über puerperal-eklampsie*, Vogel.

Schnorrenberg, F., Pattichis, C. S., Kyriacou, K. C. and Schizas, C. N.: 1997, Computer-aided detection of breast cancer nuclei, *Information Technology in Biomedicine, IEEE Transactions on* **1**(2), 128–140.

Sheikhi, G. and Altınçay, H.: 2016, The cost of type ii diabetes mellitus: A machine learning perspective, *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, Springer, pp. 818–821.

Shiu, Y., Jeong, H. and Kuo, C. J.: 2005, Musical structure analysis using similarity matrix and dynamic programming.

Snijders, R. J., Sundberg, K., Holzgreve, W., Henry, G. and Nicolaides, K. H.: 1999, Maternal age-and gestation-specific risk for trisomy 21, *Ultrasound in obstetrics & gynecology* **13**(3), 167–170.

Snijders, R., Noble, P., Sebire, N., Souka, A., Nicolaides, K. et al.: 1998, Uk multicentre project on assessment of risk of trisomy 21 by maternal age and fetal nuchal-translucency thickness at 10–14 weeks of gestation, *The Lancet* **352**(9125), 343–346.

Song, J., Bae, S.-Y. and Yoon, K.: 2002, Mid-level music melody representation of polyphonic audio for query-by-humming system., *ISMIR*.

Spencer, K.: 2007, Aneuploidy screening in the first trimester, *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, Vol. 145, Wiley Online Library, pp. 18–32.

Spencer, K., Souter, V., Tul, N., Snijders, R. and Nicolaides, K.: 1999, A screening program for trisomy 21 at 10–14 weeks using fetal nuchal translucency, maternal serum free β-human chorionic gonadotropin and pregnancy-associated plasma protein-a, *Ultrasound in obstetrics & gynecology* **13**(4), 231–237.

Spencer, K., Spencer, C. E., Power, M., Dawson, C. and Nicolaides, K. H.: 2003, Screening for chromosomal abnormalities in the first trimester using ultrasound and maternal serum biochemistry in a one-stop clinic: a review of three years prospective experience, *BJOG: An International Journal of Obstetrics & Gynaecology* **110**(3), 281–286.

Spencer, K., Tul, N., Nicolaides, K. H. et al.: 2000, Maternal serum free b-hcg and papp-a in fetal sex chromosome defects in the first trimester, *Prenat Diagn* **20**(5), 390–4.

Stefanowski, J. and Wilk, S.: 2008, Selective pre-processing of imbalanced data for improving classification performance, *International Conference on Data Warehousing and Knowledge Discovery*, Springer, pp. 283–292.

Stephan, C., Cammann, H., Meyer, H.-A., Müller, C., Deger, S., Lein, M. and Jung, K.: 2008, An artificial neural network for five different assay systems of prostate-specific antigen in prostate cancer diagnostics, *BJU international* **102**(7), 799–805.

Sundberg, K., Bang, J., Smidt-Jensen, S., Brocks, V., Lundsteen, C., Parner, J., Keiding, N. and Philip, J.: 1997, Randomised study of risk of fetal loss related to early amniocentesis versus chorionic villus sampling, *The Lancet* **350**(9079), 697–703.

Tabor, A. and Alfirevic, Z.: 2009, Update on procedure-related risks for prenatal diagnosis techniques, *Fetal diagnosis and therapy* **27**(1), 1–7.

Talkin, D.: 1995, A robust algorithm for pitch tracking, *Speech Coding and Synthesis, Edited by W. B. Klein, K. K. Palival* .

Taylor, E. R.: 1991, *The AB guide to music theory*, Associated Board of the Royal Schools of Music.

Thorndike, R. L.: 1953, Who belongs in the family?, *Psychometrika* **18**(4), 267–276.

Tzanetakis, G. and Cook, P.: 2002, Musical genre classification of audio signals, *IEEE Transactions on Speech and Audio Processing* **10**, 293 – 302.

Tzanetakis, G. and Essl, G.: 2001, Automatic musical genre classification of audio signals, *IEEE Transactions on Speech and Audio Processing* pp. 293–302.

van Kranenburg, P. and Tzanetakis, G.: 2010, A computational approach to the modeling and employment of cognitive units of folk song melodies using audio recordings, *Proceedings of the 11th International Conference on Music Perception and Cognition*, pp. 794–797.

Velarde, G., Weyde, T. and Meredith, D.: 2013, An approach to melodic segmentation and classification based on filtering with the haar-wavelet, *Journal of New Music Research* **42**(4), 325–345.

Volk, A. and Van Kranenburg, P.: 2012, Melodic similarity among folk songs: An annotation study on similarity-based categorization in music, *Musicae Scientiae* p. 1029864912448329.

Wang, Y., Li, J., Gu, J., Zhou, Z. and Wang, Z.: 2015, Artificial neural networks for infectious diarrhea prediction using meteorological factors in shanghai (china), *Applied Soft Computing* **35**, 280–290.

Werbos, P.: 1974, Beyond regression: New tools for prediction and analysis in the behavioral sciences.

Wilcoxon, F.: 1945, Individual comparisons by ranking methods, *Biometrics bulletin* **1**(6), 80–83.

Wilk, S., Stefanowski, J., Wojciechowski, S., Farion, K. J. and Michalowski, W.: 2016, Application of preprocessing methods to imbalanced clinical data: An experimental study, *Information Technologies in Medicine*, Springer, pp. 503–515.

Wolf, C. G., Koved, L. and Kunzinger, E.: 2016, * ibm tj watson research center, yorktown heights, ny 10598 usa** ibm intelligent communications services, boca raton, fl 33429 usa cwolf (a) watson. ibm. com, *Human-Computer Interaction: Interact'95* p. 247.

Wright, D., Kagan, K., Molina, F., Gazzoni, A. and Nicolaides, K.: 2008, A mixture model of nuchal translucency thickness in screening for chromosomal defects, *Ultrasound in Obstetrics & Gynecology* **31**(4), 376–383.

Wu, Z. and Li, H.: 2013, Voice conversion and spoofing attack on speaker verification systems, *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, IEEE, pp. 1–9.

Yoo, T. K., Kim, D. W., Choi, S. B., Oh, E. and Park, J. S.: 2016, Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: A cross-sectional study, *PloS one* **11**(2), e0148724.

Zade, D. A., Ardil, C. and Rustamov, S.: 2006, Investigation of combined use of mfcc and lpc features in speech recognition systems, *International Journal of Siganl Processing* **3**, 105–111.

Zelig, C. M., Knutzen, D. M., Ennen, C. S., Dolinsky, B. M. and Napolitano, P. G.: 2016, Chorionic villus sampling, early amniocentesis, and termination of pregnancy without diagnostic testing: Comparison of fetal risk following positive non-invasive prenatal testing, *Journal of Obstetrics and Gynaecology Canada* .

Zhang, H., Gao, Y., Jiang, F., Fu, M., Yuan, Y., Guo, Y., Zhu, Z., Lin, M., Liu, Q., Tian, Z. et al.: 2015, Non-invasive prenatal testing for trisomies 21, 18 and 13: clinical experience from 146 958 pregnancies, *Ultrasound in Obstetrics & Gynecology* **45**(5), 530–538.

Zheng, F., Zhang, G. and Song, Z.: 2001, Comparison of different implementations of mfcc, *J. Comput. Sci. Technol.* **16**, 582–589.

Zhu, Y. and Shasha, D.: 2003, Query by humming: a time series database approach, *Proc. of SIGMOD*, p. 675.

Research Activities

8.1 Journal Papers

- A. Neocleous, K. Nicolaides, and C. Schizas, "First Trimester Non-invasive Prenatal Diagnosis: A Computational Intelligence Approach," IEEE Journal of Biomedical and Health Informatics, Vol. 20, 2015.

- A. Neocleous, K. Nicolaides, and C. Schizas, "Intelligent Non-invasive Diagnosis of Aneuploidy: Raw Values and Highly Imbalanced Dataset," accepted for publication in IEEE Journal of Biomedical and Health Informatics, 2016.

- A. Neocleous, A. Syngelaki, K. Nicolaides, and C. Schizas, "Two Stage Approach for Aneuploidy Risk Estimation Using Computational Intelligence," submitted for publication in Ultrasound in Obstetrics and Gynecology, 2016.

- A. Neocleous, G. Azzopardi, C. N. Schizas and Nicolai Petkov, "COSFIRE flters for 1-D pattern recognition with application to musical signals" prepared for submission in journal of signal processing and music, EURASIP, 2016.

8.2 Conference and Workshop Papers

- A. Neocleous, S. Kouzoupis and I. Athanassakis, "Categorization of the experimental mice ultrasonic vocalizations during the male-female intercourse", HELINA's 5th National conference, 2010.

- A. Neocleous, R. Ramirez and A. Perez, "Modeling Emotions in Violin Audio Recordings", 3rd International Workshop on Machine Learning and Music, 2010.

- C.N. Schizas, K. Nicolaides, K. Neokleous, A. Neocleous, C. Neocleous and N. Schiza, "Computational Intelligent Diagnostic System in Predicting Chromosomal Abnormalities of the Fetus", 10th World Congress in Fetal Medicine Malta, 2011.
- C.N. Schizas, K. Nicolaides, K. Neokleous, A. Neocleous, C. Neocleous and N. Schiza, "Computational Intelligent Diagnostic System in Predicting Preeclampsia for Pregnant Women", 10th World Congress in Fetal Medicine Malta, 2011.
- C. Neocleous, K. Nicolaides, K. Neokleous, C.N. Schizas and A. Neocleous, "Artificial Neural Networks to Investigate the Significance of PAPP-A and b-hCG for the Prediction of Chromosomal Abnormalities", International Joint Conference on Neural Networks, 2011.
- A. Neocleous, K. Nicolaides A. Syngelaki, C.N. Schizas, K. Neokleous, G. Loizou and K. Neocleous, "Artificial Neural Networks to Investigate the Importance and the Sensitivity of Various Parameters used for the Prediction of Chromosomal Abnormalities", 1st Artificial Intelligence Applications in Biomedicine Workshop, 2012.
- A. Neocleous, S. Kouzoupis and I. Athanassakis, "Κατηγοριοποίηση των Υπερηχητικών Μηνυμάτων των Πειραματικών Ποντικιών Χρησιμοποιώντας Τεχνικές Μηχανικής Μάθησης ', HELINA's 5th National conference, 2012.
- A. Neocleous, M. Panteli, N. Petkov and C.N. Schizas, "Identification of Similarities between the Turkish Makam Scales and the Cypriot Folk Music", HELINA's 5th National conference, 2012.
- A. Neocleous, M. Panteli, N. Petkov and C.N. Schizas, "Timbre and Tonal Similarities Between the Turkish, Western and Cypriot Monophonic Songs Using Machine Learning Techniques", 3rd international workshop on Folk Music Analysis, Amsterdam, Netherlands, 2013.
- A. Neocleous, N. Petkov and C.N. Schizas, "Finding repeating stanzas in monophonic folk songs of Cyprus", 6th Cyprus workshop on signal processing and informatics, 2013.
- A. Neocleous, M. Panteli, R. Ioannou, N. Petkov, C.N. Schizas, "A Machine Learning Approach for Clustering Western and Non-Western Folk Music Using Low-level and Mid-level Features", 6th International Workshop on Machine Learning and Music, Prague, Czech Republic, 2013.
- A. Neocleous, N. Petkov, C.N. Schizas, "Automated Segmentation of Folk Songs Using Artificial Neural Networks", 6th International Conference in Neural Computation Theory and Applications, Rome 2014.

- A. Neocleous, N. Petkov, C. Schizas, "Automated Classification in Vocal/Instrumental parts of Folk Songs", 7^{nth} Cyprus workshop on signal processing and informatics, 2014.
- A. Neocleous, G. Azzopardi, C.N. Schizas, N. Petkov "Filter-Based Approach for Ornamentation Detection and Recognition in Singing Folk Music", Computer Analysis of Images and Patterns CAIP, Lecture Notes in Computer Science, Vol. 9256, pp. 558-569, 2015.
- A. Neocleous, C. Neocleous, N. Petkov, K. Nicolaides and C.N. Schizas, "Interpretation of the receiver operating characteristic curve for aneuploidy prenatal diagnosis", MEDICON 2016.

8.3 Participation in workshops

- Ηλεκτρονικός φάκελος ασθενή, Είναι Μονόδρομος, Πανεπιστήμιο Κύπρου, 2013.
- Ηλεκτρονική Υγεία, Αναγκαιότητα για Επιτυχημένη Εφαρμογή ενός Γενικού Σχεδίου υγείας (ΓΕΣΥ), Πανεπιστήμιο Κύπρου, 2014.

8.4 Volunteering in Conferences

- eHealth Week, Riga, Latvia, 2015
- eHealth Week, Amsterdam, The Netherlands, 2016

Curriculum Vitae

Andreas Neocleous was born on the 18^{th} of July 1984 in Larnaca, Cyprus. He studied audio signal processing and acoustics at the Technical University of Crete and he obtained a MSc degree in Sound and Music Computing from the University of Pompeu Fabra, Barcelona, Spain in 2010. In 2011, he joined the group of Intelligent Systems of the Computer Science Department of the University of Cyprus (UCY) and he worked as research scientist for medical applications in collaboration with the Fetal Medicine Foundation (FMF). In the period between 2012 and 2014 he worked in a research project funded by the EU, the UCY, and the Cyprus Research Promotion Foundation, for the computational analysis of the non-Western traditional music in the area of Middle East and he continued his collaboration with the FMF. In 2012, he started collaborating with the Intelligent Systems group at the University of Groningen as a PhD candidate. He has been funded by the University of Groningen through an "Ubbo Emmius" scholarship for international sandwich PhD programs between the period 2014 and 2016. His research interests focus on Digital Signal Processing, Machine Learning and Computational Intelligence applied in medical diagnostic systems and in ethnomusicology.